

Notes on Estimation Theory

Gyubeom Edward Im*
(Orig. by Steven Kay and Dan Simon)

February 19, 2024

Contents

1	Introduction	3
1.1	The Mathematical Estimation Problem	3
1.2	Assessing Estimator Performance	4
2	Minimum Variance Unbiased Estimation	6
2.1	Unbiased Estimators	6
2.1.1	Example 2.1 and Example 2.2	6
2.2	Minimum Variance Criterion	7
2.3	Existence of the Minimum Variance Unbiased Estimator	8
2.4	Finding the Minimum Variance Unbiased Estimator	8
3	Cramer-Rao Lower Bound	8
3.1	Estimator Accuracy Considerations	8
3.1.1	Example 3.1 - PDF Dependence on Unknown Parameter	9
3.2	Cramer-Rao Lower Bound	10
3.2.1	Theorem 3.1 (Cramer-Rao Lower Bound - Scalar Parameter)	10
3.2.2	Example 3.3 - DC Level in White Gaussian Noise	11
3.3	Transformation of Parameters	12
3.4	Extension to a Vector Parameter	13
3.4.1	Example 3.6 - DC Level in White Gaussian Noise (Revisited)	14
3.4.2	Example 3.7 - Line Fitting	15
3.5	Vector Parameter CRLB for Transformations	16
3.5.1	Example 3.8 - CRLB for Signal-to-Noise Ratio	16
3.6	CRLB for the General Gaussian Case	17
3.6.1	Example 3.11 - Random DC Level in WGN	17
4	Linear Models	18
4.1	Definition and Properties	18
4.1.1	Example 4.2 - Fourier Analysis	19
4.2	Extension to the Linear Model	21
5	General Minimum Variance Unbiased Estimation	22
5.1	Sufficient Statistics	22
5.1.1	Example 5.1 - Verification of a Sufficient Statistic	23
5.2	Finding Sufficient Statistics	23
5.2.1	Theorem 5.1 (Neyman-Fisher Factorization)	23
5.2.2	Example 5.2 - DC Level in WGN	24
5.2.3	Proof of the Neymann-Fisher Factorization	24
5.3	Using Sufficiency to Find the MVU Estimator	24
5.3.1	Example 5.5 - DC Level in WGN	24
5.3.2	Theorem 5.2 (Rao-Blackwell-Lehmann-Scheffe)	26
5.3.3	Example 5.6 - Completeness of a Sufficient Statistic	26
5.3.4	Example 5.7 - Incomplete Sufficient Statistic	27
5.4	Extension to a Vector Parameter	29
5.4.1	Example 5.10 and Example 5.11 - DC Level in WGN with Unknown Noise Power	29

*blog: alida.tistory.com, email: gyurse@gmail.com

6 Best Linear Unbiased Estimation	30
6.1 Definition of the BLUE	30
6.2 Finding the BLUE	31
6.2.1 Example 6.1 - DC Level in White Noise	33
6.3 Extension to a Vector Parameter	33
6.3.1 Theorem 6.1 (Gauss-Markov Theorem)	34
7 Maximum Likelihood Estimation	35
7.1 An Example	35
7.1.1 Example 7.1 - DC Level in White Gaussian Noise - Modified	35
7.2 Finding the MLE	37
7.2.1 Example 7.2 and 7.3 - DC Level in White Gaussian Noise - Modified (continued)	37
7.3 Properties of the MLE	38
7.3.1 Theorem 7.1 (Asymptotic Properties of the MLE)	38
8 Least Squares	38
8.1 The Least Squares Approach	38
8.1.1 Example 8.1 - DC Level Signal	39
8.2 Linear Least Squares	39
8.2.1 Scalar Case	39
8.2.2 Vector Case	40
8.2.3 Vector Weighted Case	41
8.3 Geometrical Interpretations	41
8.4 Sequential Least Squares	43
8.4.1 Sequential LS (scalar parameter)	43
8.4.2 WLS → Sequential LS	44
8.4.3 WLS → Sequential LS (vector parameter)	46
8.5 Constrained Least Squares	47
8.6 Nonlinear Least Squares	48
8.6.1 General approach for nonlinear LS	49
8.6.2 Newton-Raphson iteration	49
8.6.3 Gauss-Newton iteration	50
9 Method of Moments	51
10 The Bayesian Philosophy	51
10.1 Prior Knowledge and Estimation	51
10.2 Choosing a Prior PDF	54
10.2.1 Example 10.1 - DC Level in WGN - Gaussian Prior PDF	55
10.3 Properties of the Gaussian PDF	57
10.3.1 Theorem 10.1 (Conditional PDF of Bivariate Gaussian)	58
10.3.2 Theorem 10.2 (Conditional PDF of Multivariate Gaussian)	59
10.4 Bayesian Linear Model	60
10.4.1 Theorem 10.3 (Posterior PDF for the Bayesian General Linear Model)	61
10.5 Bayesian Estimation for Deterministic Parameters	61
11 General Bayesian Estimators	62
11.1 Risk Functions	62
11.2 Maximum A Posteriori Estimators	64
11.2.1 Example 11.2 - Exponential PDF	65
11.2.2 Scalar MAP estimator v.s. vector MAP estimator	65
11.2.3 Example 11.5 - Exponential PDF	66
11.3 Performance Description	67
11.3.1 Example 11.6 - DC Level in WGN - Gaussian Prior PDF	68
12 Linear Bayesian Estimators	68
12.1 Linear MMSE Estimation	68
12.1.1 Example 12.1 - DC Level in WGN with Uniform Prior PDF	70
12.2 Geometrical Interpretations	71
12.3 The Vector LMMSE Estimator	72
12.3.1 Theorem 12.1 (Bayesian Gauss-Markov Theorem)	73

13 Kalman Filters	74
14 References	74
15 Revision log	74

1 Introduction

추정 이론(estimation theory)은 관측된 데이터를 바탕으로 모델의 파라미터나 상태를 예측하는 다양한 방법을 정리한 이론이다. 이는 데이터 분석, 신호처리, 기계학습, 금융, 로봇공학 등 다양한 분야에서 널리 쓰이고 있으며 주로 불확실성을 다루는 과정에서 정확한 결정을 내리기 위한 필수적인 도구로 사용된다. 추정 이론의 응용 분야는 매우 넓은데 통신에서는 신호의 품질을 추정하거나 기계학습에서는 데이터로부터 알고리즘의 파라미터를 결정하는데 사용된다. 또한 금융 분야에서는 시장의 미래 동향을 예측하기 위한 변수를 추정하는데 필수적으로 사용되고 있다.

1.1 The Mathematical Estimation Problem

좋은 추정값(estimator)을 얻기 위해서는 우선 수학적으로 데이터를 잘 모델링해야 한다. 데이터는 랜덤성을 띠기 때문에 확률 밀도 함수(probability density function, pdf) $p(x[0], x[1], \dots, x[N-1]; \theta)$ 를 사용하여 데이터를 표현한다. 이 때 $x[n]$ 은 N 개의 데이터를 의미하며 θ 는 미지의 모델 파라미터를 의미한다. 만약 $N = 1$ 인 경우 pdf는 아래와 같이 나타낼 수 있다.

$$p(x[0]; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2}(x[0] - \theta)^2 \right] \quad (1)$$

- $p(x; \theta)$: 확률 분포가 파라미터 θ 에 의해 정의됨. (pdf of x parameterized by θ)

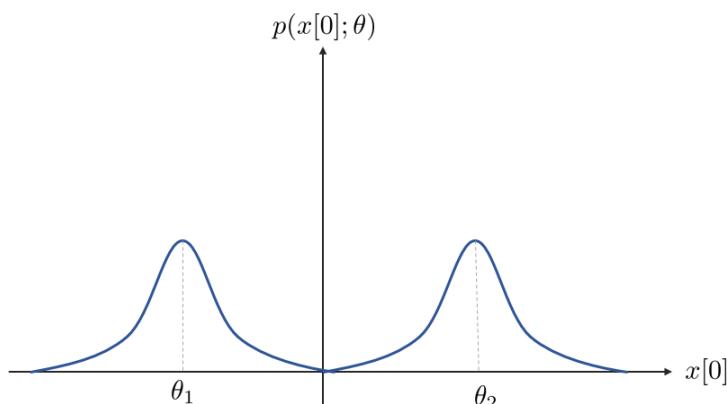
Tip

pdf를 표기하는 서로 다른 방법을 정리하면 다음과 같다.

1. $p(x; \theta)$: 매개 변수 θ 에 대한 x 의 확률 분포 (pdf of x parameterized by θ)
2. $p(x|\theta)$: θ 가 주어졌을 때 x 의 조건부 확률 분포 (pdf of x given θ)
3. $p(x, \theta)$: x, θ 가 동시에 발생할 결합 확률 분포 (joint pdf of x, θ)

1)에서 θ 는 확률 변수(random variable)일 수도 있고 아닐 수도 있다. x 는 확률 변수이다. 2), 3)에서 x 와 θ 는 반드시 확률 변수이어야 한다. 3)에서 두 확률 변수 x, θ 는 일반적으로 서로 독립이 아니며 종속적이다.

위 식에서 보다시피 파라미터 θ 는 $x[0]$ 의 확률에 직접적인 영향을 주기 때문에, 역으로 $x[0]$ 의 값을 보고 θ 를 추정하는 것이 가능하다.



예를 들면 아래와 같은 그림이 주어졌다고 했을 때 $x[0]$ 값이 음수가 관측되면 우리는 $\theta = \theta_2$ 보다는 $\theta = \theta_1$ 이라고 일반적으로 유추할 수 있다. 하지만 실제 문제에서는 위와 같은 pdf는 주어지지 않기 때문에 데이터 x 와

파라미터 θ 의 관계를 정의해야 한다. 랜덤한 노이즈를 $w[n]$ 라고 했을 때 둘 사이의 관계는 아래와 같이 정의할 수 있다.

$$x[n] = A + Bn + w[n] \quad n = 0, 1, \dots, N-1 \quad (2)$$

일반적으로 $w[n]$ 는 평균이 0인 white gaussian noise(WGN)으로 설정한다. 이 때, $\theta = [A, B]^\top$ 는 미지의 모델 파라미터를 말한다.

$$w[n] \sim \mathcal{N}(0, \sigma^2) \quad (3)$$

- $a \sim \mathcal{N}(\mu, \sigma^2)$: 확률변수 a 가 평균이 μ 이고 분산이 σ^2 인 가우시안 분포를 따른다.

N 개의 데이터를 벡터화하여 $\mathbf{x} = [x[0], x[1], \dots, x[N-1]]^\top$ 라고 표현하면 \mathbf{x} 에 대한 pdf는 다음과 같다.

$$p(\mathbf{x}; \theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} (x[n] - A - Bn)^2 \right] \quad (4)$$

pdf에 기반한 (4)과 같은 추정은 고전적인 추정 방법으로써 **파라미터 θ 를 우리가 모르지만 고정된 상수로 보는 빈도주의(frequentist) 관점**으로 해석될 수 있다. 이와 달리 현대적인 추정 방법은 **파라미터 θ 또한 별도의 확률 변수로 해석하는 베이지안(bayesian) 관점**을 주로 사용한다.

$$\begin{aligned} \text{Frequentist: } & \underbrace{x[n]}_{\text{r.v.}} = \underbrace{\theta}_{\text{deterministic}} + w[n] \\ \text{Bayesian: } & \underbrace{x[n]}_{\text{r.v.}} = \underbrace{\theta}_{\text{r.v.}} + w[n] \end{aligned} \quad (5)$$

베이지안 관점에서는 데이터 \mathbf{x} 와 파라미터 θ 가 둘 다 확률변수이므로 둘 사이의 결합 확률 분포(joint pdf)을 사용하여 확률을 표현한다.

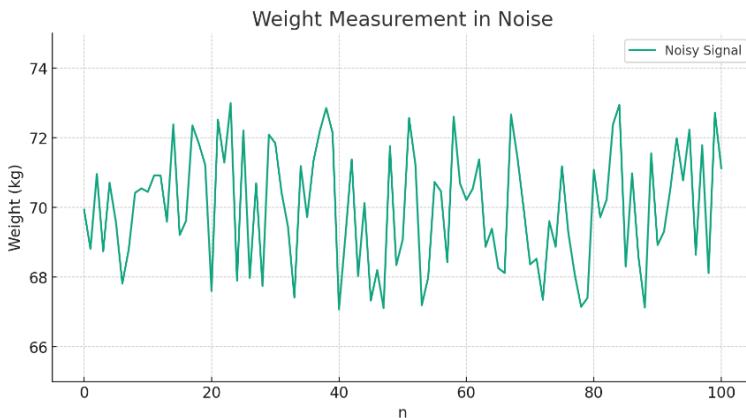
$$p(\mathbf{x}, \theta) = p(\mathbf{x}|\theta)p(\theta) \quad (6)$$

- $p(\mathbf{x}, \theta)$: joint pdf

- $p(\mathbf{x}|\theta)$: conditional pdf : θ 를 알고 있는 상태에서 데이터 \mathbf{x} 에 대한 우리의 지식

- $p(\theta)$: prior pdf: 어떤 데이터 \mathbf{x} 가 관측되기 전 θ 에 대한 우리의 경험, 지식

1.2 Assessing Estimator Performance



위와 같이 100일간 측정한 노이즈를 포함한 몸무게 데이터가 주어졌다고 가정하자. 위 데이터는 아래와 같이 모델링할 수 있다.

$$x[n] = A + w[n] \quad (7)$$

- $w[n]$: 평균이 0인 노이즈

- $x[n]$: 측정된 데이터

- A : 추정하고자 하는 파라미터

일반적으로 A 를 다음과 같이 데이터들의 평균으로 추정하는 것이 합리적일 것이다.

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad (8)$$

- \hat{A} : A 의 추정값 1

여기에서 다음과 같은 질문을 할 수 있다.

- \hat{A} 는 실제 A 와 얼마나 가까울까?
- 평균말고 더 좋은 추정 방법은 없을까?

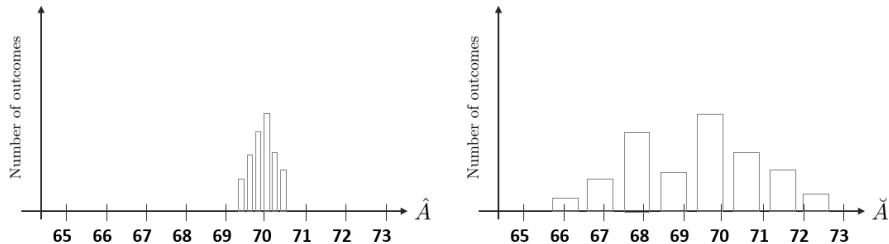
다음과 같이 다른 추정 방법을 사용하여 A 를 추정할 수 있다.

$$\check{A} = x[0] \quad (9)$$

- \check{A} : A 의 추정값 2

직관적으로 우리는 모든 데이터(또는 무한개의 데이터)를 관측하는게 아닌 이상 좋은 추정값을 얻는 것이 어렵다는 것을 알 수 있다. 실제로 $\check{A} = 69.8$ 이고 $\hat{A} = 71.1$ 이어서 \check{A} 가 $A = 70$ 에 더 가깝다. 이런 경우 \check{A} 가 더 좋은 추정값이라고 볼 수 있을까? 당연히 아니다.

추정값(estimator) \hat{A} 는 확률변수 $x[n]$ 에 대한 함수이므로 \hat{A} 역시 확률변수가 된다. 따라서 추정값 역시 노이즈로 인해 다양한 결과물들을 도출할 수 있다. \hat{A} 가 A 에 더 가깝다는 사실은 주어진 $x[n]$ 의 예제에 대해서만을 의미한다. 따라서 추정값의 성능을 평가하기 위해서는 반드시 통계적으로 접근해야 한다. 예를 들어 여러번의 실험을 통해 데이터를 수집하고 이를 반복적으로 추정하는 방법이 존재한다.



데이터를 여러번 수집한 후 추정값 \hat{A}, \check{A} 의 기대값(expectation)을 계산하면 다음과 같다.

$$\begin{aligned} \mathbb{E}(\hat{A}) &= \mathbb{E}\left(\frac{1}{N} \sum_{n=0}^{N-1} x[n]\right) \\ &= \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}(x[n]) \\ &= \frac{1}{N} \sum_{n=0}^{N-1} A \\ &= A \end{aligned} \quad (10)$$

$$\begin{aligned} \mathbb{E}(\check{A}) &= \mathbb{E}(x[0]) \\ &= A \end{aligned} \quad (11)$$

따라서 두 추정값의 성능은 동일한 것일까? \hat{A} 가 \check{A} 보다 더 좋은 추정값임을 증명하기 위해서는 추정의 분산이 더 작음을 입증해야 한다.

$$\begin{aligned} \text{var}(\hat{A}) &= \text{var}\left(\frac{1}{N} \sum_{n=0}^{N-1} x[n]\right) \\ &= \frac{1}{N^2} \sum_{n=0}^{N-1} \text{var}(x[n]) \\ &= \frac{1}{N^2} N \sigma^2 \\ &= \frac{\sigma^2}{N} \end{aligned} \quad (12)$$

$$\begin{aligned}\text{var}(\check{A}) &= \text{var}(x[0]) \\ &= \sigma^2 \\ &> \text{var}(\hat{A})\end{aligned}\tag{13}$$

따라서 이를 통해 얻을 수 있는 결론은 다음과 같다.

1. 추정값(estimator)는 확률변수(random variable)이다. 따라서 추정값의 성능은 반드시 통계적 방법이나 pdf를 통해 판단되어야 한다.
2. 컴퓨터 시뮬레이션을 사용하여 추정값을 평가하는 방법은 파라미터에 대한 통찰을 얻기엔 충분히 좋지만 이를 절대적인 값으로 해석하면 안된다. 운이 좋은 경우 추정값은 소수점 오차를 가진 정확도로 구할 수 있지만 운이 나쁜 경우에는(데이터가 부족하거나 여러 값이 들어 있거나) 잘못된 추정값을 얻을 수 있다.

2 Minimum Variance Unbiased Estimation

본 섹션에서 나오는 추정값들은 고전적인 빈도주의 관점에서 파라미터 θ 가 고정된 값으로 주어졌다고 가정한다.

2.1 Unbiased Estimators

추정값이 불편성(unbiased)을 지닌다는 의미는 추정값의 평균이 파라미터의 참 값과 동일하다는 의미와 동치이다. 일반적으로 파라미터는 특정 범위 $a < \theta < b$ 안에 존재하므로 불편추정값(unbiased estimator, 또는 불편추정량)이란 다음과 같이 수학적으로 정의할 수 있다.

$$\mathbb{E}(\hat{\theta}) = \theta \tag{14}$$

만약 추정값이 편향(biased)되어 있다면 $\mathbb{E}(\hat{\theta}) \neq \theta$ 이고 편향 $b(\theta)$ 은 다음과 같이 계산할 수 있다.

$$b(\theta) = \mathbb{E}(\hat{\theta}) - \theta \tag{15}$$

2.1.1 Example 2.1 and Example 2.2

다음과 같은 데이터가 주어졌다고 하자.

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1 \tag{16}$$

- A : 추정하고자 하는 파라미터, $-\infty < A < \infty$

- $w[n]$: WGN

이에 대한 일반적인 추정값 \hat{A} 는 다음과 같이 데이터의 평균으로 예측할 수 있다.

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \tag{17}$$

선형성에 의해 기대값(expectation)은 다음과 같이 정의된다.

$$\begin{aligned}\mathbb{E}(\hat{A}) &= \mathbb{E}\left(\frac{1}{N} \sum_{n=0}^{N-1} x[n]\right) \\ &= \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}(x[n]) \\ &= \frac{1}{N} \sum_{n=0}^{N-1} A \\ &= A\end{aligned}\tag{18}$$

따라서 평균 추정값 \hat{A} 는 불편추정값(unbiased estimator)이다. 만약 다음과 같은 추정값 \check{A} 가 있다고 가정해보자.

$$\check{A} = \frac{1}{2N} \sum_{n=0}^{N-1} x[n] \tag{19}$$

\check{A} 의 기대값은 다음과 같다.

$$\begin{aligned}\mathbb{E}(\check{A}) &= \frac{1}{2}A \\ &= A \text{ if } A = 0 \\ &\neq A \text{ if } A \neq 0\end{aligned}\tag{20}$$

따라서 \check{A} 는 편의추정값(biased estimator)이다.

불편추정값이 반드시 좋은 추정값을 의미할까? 추정값이 불편성을 지닌다고 해서 반드시 좋은 추정값이라는 의미는 아니다. 불편성의 의미는 오직 추정값의 기대값(expectation)이 실제 값과 동일하다는 의미만 지닌다. 이와 반대로 편의추정값은 시스템의 노이즈를 포함하여 모델링한 값일 수 있다. 하지만 영구적인 편향성은 항상 안 좋은 추정값을 가진다. 예를 들면 다음과 같이 동일 파라미터 θ 에 대한 여러 추정값 $\{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n\}$ 이 주어졌을 때 가장 합리적인 방법은 이들의 기대값을 구하는 것이다.

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i\tag{21}$$

만약 모든 추정값들이 불편성을 지니고 서로 독립이라면 다음 공식이 성립한다.

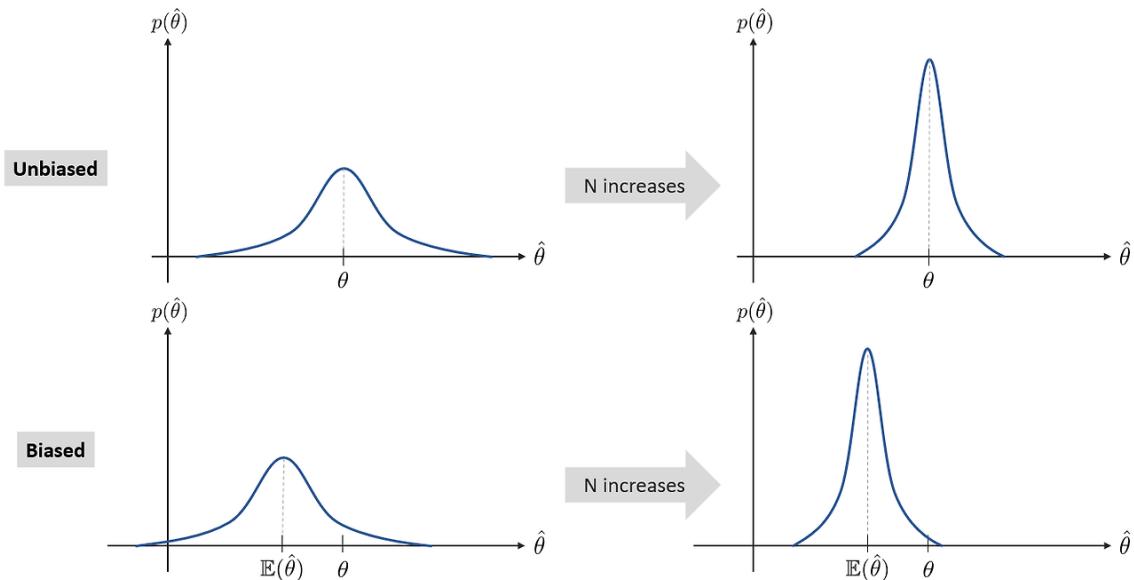
$$\mathbb{E}(\hat{\theta}) = \theta\tag{22}$$

$$\begin{aligned}\text{var}(\hat{\theta}) &= \frac{1}{n^2} \sum_{i=1}^n \text{var}(\hat{\theta}_i) \\ &= \frac{\text{var}(\hat{\theta}_1)}{n}\end{aligned}\tag{23}$$

따라서 위 식에서 보다시피 많은 수($n \uparrow$)의 추정값을 사용할 수록 분산은 작아진다. 만약 $n \rightarrow \infty$ 이면 분산은 0이 되고 $\hat{\theta} \rightarrow \theta$ 가 된다. 하지만 편의추정량의 경우 $\mathbb{E}(\hat{\theta}_i) = \theta + b(\theta)$ 이므로 다음과 같은 기대값을 가진다.

$$\begin{aligned}\mathbb{E}(\hat{\theta}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\hat{\theta}_i) \\ &= \theta + b(\theta)\end{aligned}\tag{24}$$

따라서 n 이 충분히 많다고 하더라도 편향 $b(\theta)$ 값은 제거되지 않으므로 실제 추정값으로 수렴하지 않는다.



2.2 Minimum Variance Criterion

최적의 추정값을 찾기 위해서는 최적의 criterion을 사용해야 한다. 널리 사용되는 criterion 중 하나가 mean square error(MSE)이다.

$$\boxed{\text{mse}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]} \quad (25)$$

- θ : 추정해야 할 파라미터. 빈도주의 관점에 의해 θ 는 고정된 상수 값을 의미한다. 즉, 확률변수가 아니다.

MSE는 실제 값 θ 과 추정값 $\hat{\theta}$ 의 평균 제곱 편차를 측정한다. MSE는 널리 사용되는 criterion 중 하나이지만 어렵게도 편향에 의한 영향을 받는다. 위 식에 $\pm \mathbb{E}(\hat{\theta})$ 를 추가한 후 식을 전개하면 다음과 같다.

$$\begin{aligned} \text{mse}(\hat{\theta}) &= \mathbb{E}\left\{\left[\left(\hat{\theta} - \mathbb{E}(\hat{\theta})\right) + \left(\mathbb{E}(\hat{\theta}) - \theta\right)\right]^2\right\} \\ &= \text{var}(\hat{\theta}) + \left[\mathbb{E}(\hat{\theta}) - \theta\right]^2 \\ &= \text{var}(\hat{\theta}) + b^2(\theta) \end{aligned} \quad (26)$$

따라서 최적의 추정값을 찾기 위한 criterion을 고려할 때 MSE를 최소화해주는 minimum MSE(MMSE) 추정값을 고려하면 안된다. MMSE에 대한 대안으로 편향이 0이고 분산을 최소화하는 추정값을 사용해야 하는데 이를 minimum variance unbiased estimator(MVUE)라고 한다. 불편추정값의 분산을 최소화하는 과정은 pdf 관점에서 $p(\hat{\theta} - \theta)$ 를 0 주변에 집중시키는 효과가 있다. 따라서 추정 오차가 커질 가능성이 작아진다.

2.3 Existence of the Minimum Variance Unbiased Estimator

독자는 MVUE가 실제로 존재하는지 여부에 대해 궁금증이 생길 수 있다. 즉, 모든 파라미터 θ 에 대해 최소의 분산을 가지며 불편향된 추정값이 존재하는지 궁금할 수 있다. 결론만 말하자면 MVUE는 항상 존재하는 것은 아니다.

2.4 Finding the Minimum Variance Unbiased Estimator

만약 MVUE가 존재한다고해도 이를 찾는 것이 불가능할 수 있다. MVUE를 찾을 수 있는 절대적인 방법이란 아직 알려지지 않았다. 하지만 이를 가능하게 해주는 몇몇 기준들은 존재한다.

1. Cramer-Rao lower bound(CRLB)를 결정하고 추정값들이 이를 만족하는지 확인한다.
2. Rao-Blackwell-Lehmann-Scheffe(RBLS) 이론을 적용한다.
3. 추정값이 불편성(unbiased) 뿐만 아니라 선형(linear) 특성이 있다는 제약조건 하에 분산을 최소화하는 MVUE를 구한다.

1,2 방법을 사용하면 MVUE를 구할 수 있고 3 방법은 MVUE가 데이터에 대하여 선형인 경우에만 적용된다.

- 1,2) CRLB는 임의의 불편추정값에 대하여 분산의 하한선(lower bound)를 결정하게 해준다. 만약 어떤 추정 값이 CRLB와 동일한 분산 값을 가진다면 이는 반드시 MVUE가 된다. CRLB와 동일한 분산 값을 가지지 않는다고 하더라도 MVUE가 존재할 수 있다. 이럴 때는 RBLS를 적용한다. RBLS는 충분통계량(sufficient statistics)을 먼저 구한 후 충분통계량에 대한 추정을 수행하는데 이 때 추정값이 θ 에 대한 불편추정값이 된다.
- 3) 이는 추정값이 선형이어야 하는 제약조건을 가진다. 이는 때때로 강력한 제약조건이지만 최적의 선형 추정값을 구할 수 있다.

3 Cramer-Rao Lower Bound

어떠한 불편추정값(unbiased estimator)의 분산의 하한선(lower bound)를 결정할 수 있다는 것은 실제 추정 문제에서 매우 유용하게 사용된다. 최선의 경우 특정 추정값이 MVUE임을 바로 구할 수도 있다. 그렇지 않은 경우라도 하더라도 여러 불편추정량에 대한 벤치마크 용도로 활용될 수도 있다. 이는 신호 추정 분야에서 매우 유용하게 활용된다. Cramer-Rao lower bound(CRLB) 이외에도 [McAulay and Hofstetter 1971, Kendall and Stuart 1979, Seidman 1970, Ziv and Zakai 1969]와 같이 분산의 한계(bound)를 결정하는 알고리즘들이 존재하지만 CRLB가 이들 중 가장 쉽게 한계를 구할 수 있다.

3.1 Estimator Accuracy Considerations

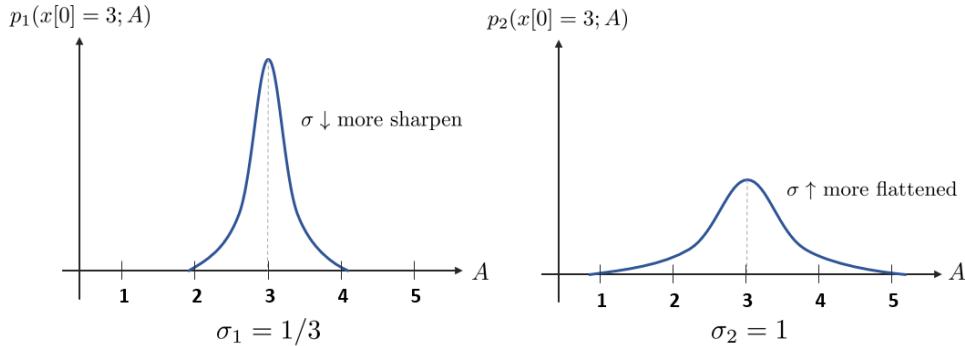
추정에 사용되는 정보는 일반적으로 관측된 데이터로부터 얻을 수 있고 관측 데이터는 노이즈를 포함하기 때문에 일반적으로 pdf로 표현될 수 있다. 따라서 추정의 정확도는 당연히 pdf와 직접적인 관련이 있다. 만약 파라미터가 pdf에 영향을 거의 주지 않는 최악의 경우에는 좋은 추정값을 얻는 것이 어려울 것이다. 따라서 파라미터가 pdf에 영향을 많이 줄수록 추정의 정확도는 올라간다.

3.1.1 Example 3.1 - PDF Dependence on Unknown Parameter

만약 하나의 데이터가 샘플링되었다고 가정해보자

$$x[0] = A + w[0] \quad (27)$$

이 때, $w[n]$ 은 $\mathcal{N}(0, \sigma^2)$ 을 따르는 white gaussian noise(WGN)이다. 일반적으로 좋은 추정값(estimator)이란 σ^2 가 작은 추정값임을 알 수 있다. 그리고 좋은 불편추정값은 $\hat{A} = x[0]$ 임을 알 수 있다.



분산 σ^2 값이 작을 수록 좋은 추정값임을 설명하기 위해 아래와 같은 예제를 들 수 있다. 만약 서로 다른 두 분산 $\sigma_1 = 1/3$ 과 $\sigma_2 = 1$ 이 주어졌고 $x[0] = 3$ 인 경우를 가정해보자. 이에 대한 pdf는 다음과 같다.

$$p_i(x[0]; A) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2}(x[0] - A)^2\right] \quad i = 1, 2 \quad (28)$$

$\sigma_1^2 < \sigma_2^2$ 이므로 우리는 $p_1(\cdot)$ 이 A 를 더 잘 추정하고 있다고 판단할 수 있다. 예를 들어 $A = 4.5$ 일 확률은 p_1 보다 p_2 가 더 높으므로 우리는 p_1 이 더 좁은 범위에 대한 확실한 확률분포를 가짐을 알 수 있다.

지금까지 설명한 pdf는 전부 데이터 x 가 주어졌을 때 파라미터 A 를 찾는 형태이기 때문에 이는 가능도함수(likelihood function) 관점에서 해석할 수 있다. **가능도함수의 뾰족한 정도(sharpness)**는 추정값이 얼마나 정확한지에 대한 정확도를 판단하는데 사용된다. 수학적으로 곡선의 뾰족한 정도는 함수의 2차 미분을 수행하여 곡률(curvature)를 구함으로써 얻을 수 있다. 이 때, pdf 특성상 exponential 항이 존재하여 계산이 어렵기 때문에 일반적으로 log를 취한 값을 사용하는데 이를 **로그 가능도함수(log likelihood function)**이라고 한다.

$$\ln p(x[0]; A) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x[0] - A)^2 \quad (30)$$

파라미터 A 에 대한 1차 미분을 수행하면 다음과 같다.

$$\frac{\partial \ln p(x[0]; A)}{\partial A} = \frac{1}{\sigma^2}(x[0] - A) \quad (31)$$

다시 한번 2차 미분을 취한 후 양변에 음수를 곱하면 다음과 같다.

$$-\frac{\partial^2 \ln p(x[0]; A)}{\partial A^2} = \frac{1}{\sigma^2} \quad (32)$$

위 식의 의미는 σ^2 가 감소할수록 원래 함수의 곡률(curvature)을 의미하는 $-\frac{\partial^2 \ln p(x[0]; A)}{\partial A^2}$ 는 증가하는 것을 의미한다. 앞서 말한 추정값 $\hat{A} = x[0]$ 의 분산은 다음과 같다.

$$\text{var}(\hat{A}) = \sigma^2 \quad (33)$$

여기에서 (32)을 대입하면 다음과 같다.

$$\text{var}(\hat{A}) = \frac{1}{-\frac{\partial^2 \ln p(x[0]; A)}{\partial A^2}} \quad (34)$$

위와 같은 간단한 예제에서는 로그 가능도함수의 2차 미분값이 데이터 $x[0]$ 에 독립이지만 일반적으로는 2차 미분값은 데이터 $\mathbf{x} = [x[0], \dots, x[n]]$ 에 종속적이다. 곡률의 정확한 표현은 다음과 같이 나타낼 수 있다.

$$\boxed{-\mathbb{E}\left[\frac{\partial^2 \ln p(x[0]; A)}{\partial A^2}\right]} \quad (35)$$

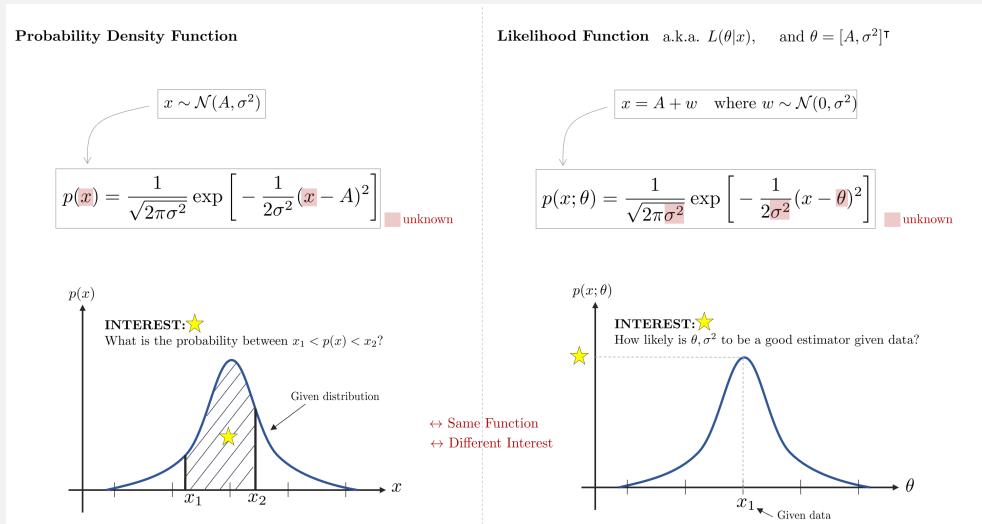
위 식을 통해 다양한 관측 데이터가 주어졌을 때 로그 가능도함수의 평균적인 곡률을 측정할 수 있다.

Tip

Likelihood Function v.s. Probability Density Function

만약 pdf $p(x; \theta)$ 가 x 는 고정된 값이면서 동시에 파라미터 θ 에 대한 함수라면 이를 일반적으로 가능도함수(likelihood function)라고 부른다. 이와 반대로, $p(x; \theta)$ 가 θ 는 고정된 값이면서 동시에 x 에 대한 함수라면 이는 일반적인 확률밀도함수(probability density function, pdf)라고 부른다. 두 개념을 비교한 그림은 아래와 같다. $p(x; \theta)$ 는 다음과 같다.

$$p(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2}(x - \theta)^2 \right] \quad (29)$$



그림에서 보는 것과 같이 pdf와 가능도함수는 모양만 같을 뿐 이를 해석하는 관점이 서로 다르다. pdf의 경우 파라미터가 주어졌을 때 특정 구간 내에서 확률을 구하는 것에 관심이 있다면 가능도함수는 데이터가 주어졌을 때 이를 가장 잘 설명하는 파라미터는 무엇인가?에 관심이 있다.

3.2 Cramer-Rao Lower Bound

3.2.1 Theorem 3.1 (Cramer-Rao Lower Bound - Scalar Parameter)

pdf $p(\mathbf{x}; \theta)$ 가 다음과 같은 정규 조건(regularity condition)을 만족하면

$$\mathbb{E} \left[\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} \right] = 0 \quad \text{for all } \theta \quad (36)$$

임의의 불편추정값(unbiased estimator) $\hat{\theta}$ 의 분산은 다음 조건을 반드시 만족한다.

$$\text{var}(\hat{\theta}) \geq \frac{1}{-\mathbb{E} \left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2} \right]} \quad (37)$$

이 때 편미분은 실제 파라미터의 참 값 θ 에 대하여 수행되었다. 추가적으로 불편추정값의 분산이 하한선(lower bound)에 도달하려면 다음 조건을 반드시 만족해야 한다. (필요충분조건)

$$\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = I(\theta)(g(\mathbf{x}) - \theta) \quad (38)$$

위 식은 로그 가능도함수의 1차 미분이 위와 같은 임의의 함수 I, g 의 곱셈 형태로 표현되어야 한다는 뜻이다. 이 때 불편추정값은 $\hat{\theta} = g(\mathbf{x})$ 가 되며 $\hat{\theta}$ 는 MVUE를 만족한다. 이 때의 최소 분산 값은 $1/I(\theta)$ 이 되며 이를 Fisher information이라고 한다. Fisher information은 일반적으로 다음과 같이 정의한다.

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2} \right] \quad (39)$$

따라서 MVUE의 분산은 다음과 같이 표현하기도 한다.

$$\text{var}(\hat{\theta}) = \frac{1}{I(\theta)} \quad \dots \text{ for MVUE} \quad (40)$$

(37)에서 2차 미분값은 \mathbf{x} 에 종속적이기 때문에 기대값은 정의에 따라 다음과 같이 전개된다.

$$\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2}\right] = \int \frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2} p(\mathbf{x}; \theta) d\mathbf{x} \quad (41)$$

3.2.2 Example 3.3 - DC Level in White Gaussian Noise

다음과 같은 여러 관측 데이터들이 주어졌다고 하자.

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1 \quad (42)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

파라미터 A 에 대한 CRLB를 유도해보면 다음과 같다. 우선 모든 관측값 \mathbf{x} 에 대한 가능도함수를 구한다.

$$\begin{aligned} p(\mathbf{x}; A) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n] - A)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right] \end{aligned} \quad (43)$$

로그 가능도 함수는 다음과 같다.

$$\ln p(\mathbf{x}; A) = -\ln[(2\pi\sigma^2)^{\frac{N}{2}}] - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \quad (44)$$

1차 미분을 수행하면 다음과 같다.

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x}; A)}{\partial A} &= \frac{\partial}{\partial A} \left[-\ln[(2\pi\sigma^2)^{\frac{N}{2}}] - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \\ &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) \\ &= \frac{N}{\sigma^2} (\bar{x} - A) \end{aligned} \quad (45)$$

- \bar{x} : \mathbf{x} 의 평균

2차 미분을 수행하면 다음과 같다.

$$\frac{\partial^2 \ln p(\mathbf{x}; A)}{\partial A^2} = -\frac{N}{\sigma^2} \quad (46)$$

위 식은 파라미터가 없는 상수임에 유의한다. (37) 식에 이를 대입하면 다음과 같다.

$$\text{var}(\hat{A}) \geq \frac{\sigma^2}{N} \quad (47)$$

위 식이 정규 조건을 만족하는가? (36)를 보면 적용해보면 만족하는 것을 알 수 있다.

$$\mathbb{E}\left[\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta}\right] = \mathbb{E}\left[\frac{N}{\sigma^2}(\bar{x} - A)\right] = 0 \quad (48)$$

- $\mathbb{E}(\bar{x}) = A$

따라서 CRLB 정의에 따라 임의의 불편추정값 \check{A} 의 분산이 $\text{var}(\check{A}) = \frac{\sigma^2}{N}$ 을 만족하면 이는 반드시 MVUE 가 된다. 위 예제에서는 $\check{A} = \bar{x}$ 가 MVUE가 된다. 그리고 이러한 추정값을 **efficient**하다라고 한다. 관측 데이터를 효율적(efficient)으로 사용하여 추정하였다는 의미이다.

3.3 Transformation of Parameters

실제 추정 문제에서는 우리가 추정하고자 하는 파라미터가 단순한 θ 가 아닌 θ 의 함수 형태를 추정해야 하는 일이 자주 발생한다. 이전 예제에서도 단순히 A 를 추정하는 것이 아닌 A^2 를 추정하고 싶을 수도 있다. 만약 A 의 CRLB를 알고 있는 경우에는 A^2 의 CRLB도 쉽게 구할 수 있고 A 와 관련된 어떤 함수라도 구할 수 있다.

추정하고자 하는 파라미터가 $\alpha = g(\theta)$ 와 같이 θ 에 대한 함수일 경우 CRLB는 다음과 같다.

$$\text{var}(\hat{\alpha}) \geq \frac{\left(\frac{\partial g}{\partial \theta}\right)^2}{-\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2}\right]} \quad (49)$$

만약 $\alpha = g(A) = A^2$ 인 경우 CRLB는 다음과 같다.

$$\text{var}(\hat{\alpha}) \geq \frac{(2A)^2}{N/\sigma^2} = \frac{4A^2\sigma^2}{N} \quad (50)$$

이전 Example 3.3 예제에서 $\hat{A} = \bar{x}$ 는 A 에 대하여 efficient함을 보였다. 따라서 \bar{x}^2 역시 A^2 에 대하여 efficient하다고 예상할 수 있으나 이는 사실이 아니다. 그 이전에 $\hat{A}^2 = \bar{x}^2$ 는 A^2 에 대한 불편추정값 조차 아니다. 즉, 편향(bias)이 존재한다.

$$\begin{aligned} \mathbb{E}(\bar{x}^2) &= \mathbb{E}^2(\bar{x}) + \text{var}(\bar{x}) \\ &= A^2 + \frac{\sigma^2}{N} \\ &\neq A^2 \end{aligned} \quad (51)$$

따라서 우리는 위 예제를 통해 추정값의 efficiency는 비선형 변환에 의해 보존되지 않는 것을 알 수 있다. 하지만 linear 또는 affine 변환에 대하여는 efficiency가 보존됨을 쉽게 보일 수 있다.

만약 $\hat{\theta}$ 가 θ 에 대하여 efficient하고 $\alpha = g(\theta) = a\theta + b$ 와 같은 affine 변환이 주어진 경우를 확인해보자.

$$\hat{\alpha} = g(\hat{\theta}) = a\hat{\theta} + b \quad (52)$$

$$\begin{aligned} \mathbb{E}(a\hat{\theta} + b) &= a\mathbb{E}(\hat{\theta}) + b \\ &= a\theta + b \\ &= g(\theta) \end{aligned} \quad (53)$$

$$\begin{aligned} \text{var}(\hat{\alpha}) &= \text{var}(a\hat{\theta} + b) \\ &= a^2\text{var}(\hat{\theta}) \end{aligned} \quad (54)$$

$g(\theta)$ 의 CRLB를 보면 다음과 같다.

$$\begin{aligned} \text{var}(\hat{\alpha}) &\geq \frac{\left(\frac{\partial g}{\partial \theta}\right)^2}{I(\theta)} \\ &= \left(\frac{\partial g}{\partial \theta}\right)^2 \text{var}(\hat{\theta}) \\ &= a^2\text{var}(\hat{\theta}) \end{aligned} \quad (55)$$

위 식에서 (54), (55)의 분산이 동일하기 때문에 $\hat{\alpha}$ 역시 MVUE이면서 동시에 efficient함을 알 수 있다.

앞서 보았듯이 efficiency는 linear 또는 affine 변환에서만 보존되는 것을 확인하였다. **하지만 데이터가 충분히 큰 경우에는 비선형 변환도 근사적으로(approximately) efficiency가 보존된다.** 다시 이전 예제 $\alpha = g(A) = A^2$ 로 돌아가서 데이터 $x[n]$ 의 평균을 $\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$ 이라고 할 때 \bar{x}^2 는 N 이 충분히 클 경우 근사적으로 편향이 제거된다.

$$\begin{aligned} \text{var}(\bar{x}^2) &= \mathbb{E}(\bar{x}^4) - \mathbb{E}^2(\bar{x}^2) \\ &= \frac{4A^2\sigma^2}{N} + \frac{2\sigma^2}{N^2} \\ &\approx \frac{4A^2\sigma^2}{N} \quad \dots \text{ for } N \rightarrow \infty \end{aligned} \quad (56)$$

$g(A) = A^2$ 의 CRLB는 다음과 같다.

$$\begin{aligned}\text{var}(\hat{\alpha}) &\geq \frac{(2A)^2}{N/\sigma^2} \\ &= \frac{4A^2\sigma^2}{N}\end{aligned}\tag{57}$$

(56), (56)가 N 이 충분히 큰 경우에 대하여 서로 동일하기 때문에 데이터가 많은 경우에는 비선형 변환에 대한 추정값도 MVUE가 되며 동시에 efficient함을 알 수 있다.

다른 방법을 사용하여 비선형 변환이 근사적으로 efficient함을 보일 수 있다. 확률분포 관점에서 봤을 때 N 이 커질수록 \bar{x} 는 A 주변으로 뾰족해지는 경향이 있다. 이에 따라 $\pm 3\sigma$ 사이의 간격은 좁아지게 되고 좁은 영역에 대해 비선형 변환을 수행하면 근사적으로 선형 변환을 한 것과 유사한 효과를 얻는다. 이를 $\bar{x} = A$ 지점에서 테일러 1차 근사를 통해 표현하면 다음과 같다.

$$g(\bar{x}) \approx g(A) + \frac{dg(A)}{dA}(\bar{x} - A)\tag{58}$$

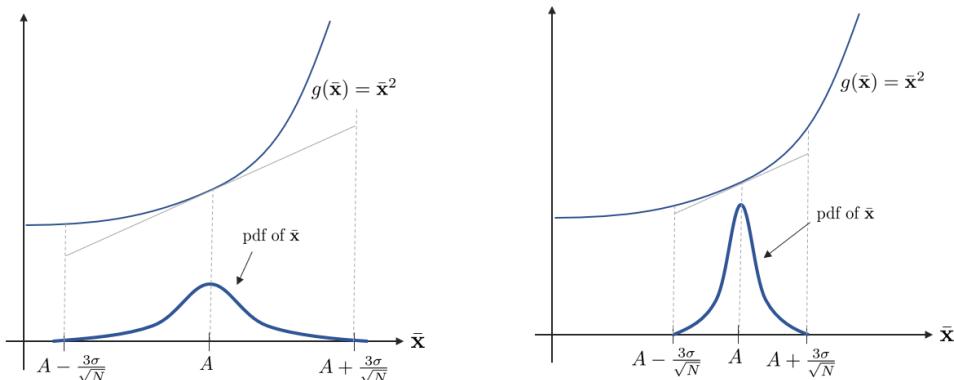
이러한 경우를 점근적으로(asymptotically) efficient하다고 한다. 이 때 기대값은 다음과 같다.

$$\mathbb{E}[g(\bar{x})] = g(A) = A^2\tag{59}$$

분산은 다음과 같다.

$$\begin{aligned}\text{var}[g(\bar{x})] &= \left[\frac{dg(A)}{dA} \right]^2 \text{var}(\bar{x}) \\ &= \frac{(2A)^2\sigma^2}{N} \\ &= \frac{4A^2\sigma^2}{N}\end{aligned}\tag{60}$$

즉 추정값은 CRLB에 점근적으로(asymptotically) 도달하는 것을 알 수 있다. 따라서 비선형 변환은 점근적으로 efficient하다.



3.4 Extension to a Vector Parameter

지금까지는 추정하려는 파라미터 θ 가 스칼라 값이었다. 해당 섹션에서는 파라미터가 여러개인 벡터 파리미터 $\theta = [\theta_1, \theta_2, \dots, \theta_p]^\top$ 케이스에 대해 다룬다. $\hat{\theta}$ 는 θ 에 대한 불편추정값(unbiased estimator)이라고 가정한다. 벡터 파라미터의 각 원소의 분산은 다음과 같이 나타낼 수 있다.

$$\text{var}(\hat{\theta}_i) \geq [\mathbf{I}^{-1}(\theta)]_{ii}\tag{61}$$

- $\mathbf{I}(\theta) \in \mathbb{R}^{p \times p}$: Fisher information 행렬

이는 (40)의 벡터 버전임을 알 수 있다. 일반적으로 스칼라 버전에서 행렬은 벡터 버전에서 역행렬로 표현된다. Fisher information 행렬을 자세히 나타내면 다음과 같다.

$$[\mathbf{I}(\theta)]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta_i \partial \theta_j}\right]\tag{62}$$

- $i = 1, 2, \dots, p$
- $j = 1, 2, \dots, p$

$p = 1$ 인 스칼라 케이스는 $\mathbf{I}(\boldsymbol{\theta}) = I(\theta)$ 가 된다. 스칼라 버전과 동일하게 불편추정값의 분산이 하한선(lower bound)에 도달하려면 다음 조건을 반드시 만족해야 한다. (필요충분조건)

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}) \quad (63)$$

위 식은 로그 가능도함수의 1차 미분이 위와 같은 임의의 함수 \mathbf{I}, \mathbf{g} 의 곱셈 형태로 표현되어야 한다는 뜻이다. 이 때 불편추정값은 $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$ 가 되며 $\hat{\boldsymbol{\theta}}$ 는 MVUE를 만족한다. 이 때의 최소 분산 값은 $1/\mathbf{I}(\boldsymbol{\theta})$ 이 된다.

$$\text{var}(\hat{\boldsymbol{\theta}}) = 1/\mathbf{I}(\boldsymbol{\theta}) \quad \dots \text{ for MVUE} \quad (64)$$

3.4.1 Example 3.6 - DC Level in White Gaussian Noise (Revisited)

예제 3.3과 같이 관측 데이터들이 주어졌다고 하자.

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1 \quad (65)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

이 때, 추정하고자 하는 파라미터가 벡터 파라미터 $\boldsymbol{\theta} = [A, \sigma^2]^T$ 인 경우를 생각해보자. 즉 $p = 2$ 이다. Fisher information 행렬은 다음과 같이 나타낼 수 있다.

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial A^2}\right] & -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial A \partial \sigma^2}\right] \\ -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \sigma^2 \partial A}\right] & -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \sigma^2 \partial \sigma^2}\right] \end{bmatrix} \in \mathbb{R}^{2 \times 2} \quad (66)$$

Fisher information 행렬은 대칭이며 동시에 positive definite한 특징을 가지고 있다. 예제 3.3의 로그 가능도함수는 다음과 같다.

$$\ln p(\mathbf{x}; \boldsymbol{\theta}) = -\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \quad (67)$$

로그 가능도함수의 1,2차 편미분은 다음과 같이 쉽게 구할 수 있다.

$$\begin{aligned} \frac{\ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial A} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) \\ \frac{\ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=0}^{N-1} (x[n] - A)^2 \\ \frac{\ln^2 p(\mathbf{x}; \boldsymbol{\theta})}{\partial A^2} &= -\frac{N}{\sigma^2} \\ \frac{\ln^2 p(\mathbf{x}; \boldsymbol{\theta})}{\partial A \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_{n=0}^{N-1} (x[n] - A) \\ \frac{\ln^2 p(\mathbf{x}; \boldsymbol{\theta})}{\partial \sigma^2 \partial \sigma^2} &= \frac{N}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{n=0}^{N-1} (x[n] - A)^2 \end{aligned} \quad (68)$$

이를 Fisher information 행렬에 대입하면 다음과 같다.

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix} \quad (69)$$

흔한 경우는 아니지만 예제의 케이스는 역행렬을 매우 쉽게 구할 수 있다. 단순히 역수를 취해줌으로써 역행렬을 구하면 (61)와 같이 CRLB를 구할 수 있다.

$$\begin{aligned} \text{var}(\hat{A}) &\geq \frac{\sigma^2}{N} \\ \text{var}(\hat{\sigma^2}) &\geq \frac{2\sigma^4}{N} \end{aligned} \quad (70)$$

이 중 $\text{var}(\hat{A})$ 는 스칼라 케이스에서 σ^2 값이 이미 주어진 경우와 동일한 CRLB를 가지는 것을 알 수 있다. 다시 말하자면 이러한 예제는 일반적인 상황에서는 참이 아니지만 예제의 경우 참이다.

3.4.2 Example 3.7 - Line Fitting

다음과 같은 line fitting 문제가 주어졌다고 가정해보자

$$x[n] = A + Bn + w[n] \quad n = 0, 1, \dots, N-1 \quad (71)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

이 때, y절편의 값 A 와 기울기 B 의 값을 찾아야 한다. 추정하고자 하는 파라미터는 $\theta = [A, B]^\top$ 이다. $p = 2$ 케이스이므로 Fisher information 행렬은 다음과 같이 나타낼 수 있다.

$$\mathbf{I}(\theta) = \begin{bmatrix} -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial A^2}\right] & -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial A \partial B}\right] \\ -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial B \partial A}\right] & -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial B^2}\right] \end{bmatrix} \in \mathbb{R}^{2 \times 2} \quad (72)$$

가능도함수는 다음과 같다.

$$p(\mathbf{x}, \theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn)^2\right] \quad (73)$$

로그 가능도함수의 1,2차 미분은 다음과 같다.

$$\begin{aligned} \frac{\ln p(\mathbf{x}; \theta)}{\partial A} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn) \\ \frac{\ln p(\mathbf{x}; \theta)}{\partial B} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn)n \\ \frac{\ln^2 p(\mathbf{x}; \theta)}{\partial A^2} &= -\frac{N}{\sigma^2} \\ \frac{\ln^2 p(\mathbf{x}; \theta)}{\partial A \partial B} &= -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} n \\ \frac{\ln^2 p(\mathbf{x}; \theta)}{\partial \sigma^2} &= -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} n^2 \end{aligned} \quad (74)$$

이를 Fisher information 행렬에 대입하면 다음과 같다.

$$\begin{aligned} \mathbf{I}(\theta) &= \frac{1}{\sigma^2} \begin{bmatrix} N & \sum_{n=0}^{N-1} n \\ \sum_{n=0}^{N-1} n & \sum_{n=0}^{N-1} n^2 \end{bmatrix} \\ &= \frac{1}{\sigma^2} \begin{bmatrix} N & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{6} \end{bmatrix} \end{aligned} \quad (75)$$

$\mathbf{I}(\theta)$ 의 역행렬을 구해보면 다음과 같다.

$$\mathbf{I}(\theta)^{-1} = \sigma^2 \begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)} \\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)} \end{bmatrix} \quad (76)$$

(61)와 같이 CRLB를 구해보면 다음과 같다.

$$\begin{aligned} \text{var}(\hat{A}) &\geq \frac{2(2N-1)\sigma^2}{N(N+1)} \\ \text{var}(\hat{B}) &\geq \frac{12\sigma^2}{N(N^2-1)} \end{aligned} \quad (77)$$

위 예제에서 보다시피 스칼라 파라미터 $\theta = A$ 만 추정했을 때와는 달리 $\theta = [A, B]^\top$ 처럼 벡터 파라미터를 추정하면 CRLB는 커지는 것을 알 수 있다. 스칼라 파라미터만 추정했을 때 CRLB는 다음과 같다.

$$\text{var}(\hat{A}) \geq \frac{\sigma^2}{N} \quad (78)$$

따라서 $N \geq 2$ 인 경우에는 벡터 파라미터의 CRLB가 항상 스칼라 파라미터의 CRLB보다 크다.

$$\frac{2(2N-1)\sigma^2}{N(N+1)} > \frac{\sigma^2}{N} \quad \dots \text{for } N \geq 2 \quad (79)$$

또한 특정 파라미터는 다른 파라미터보다 데이터 개수 N 에 민감하게 반응할 수 있다.

$$\frac{\text{CRLB}(\hat{A})}{\text{CRLB}(\hat{B})} = \frac{(2N-1)(N-1)}{6} > 1 \quad \dots \text{for } N \geq 3 \quad (80)$$

$\text{CRLB}(\hat{B})$ 는 데이터 증가에 $1/N^3$ 으로 감소하는 반면, $\text{CRLB}(\hat{A})$ 는 데이터 증가에 $1/N$ 비율로 감소한다. 이러한 차이로 인해 $x[n]$ 이 B 를 변경하는 것에 A 를 변경하는 것보다 더 민감하게 반응한다는 것을 알 수 있다.

3.5 Vector Parameter CRLB for Transformations

스칼라 파라미터의 변환에 대해 이전 섹션에서 알아본 것처럼, 실제 추정 문제에서는 단순히 벡터 파라미터 θ 를 추정해야 하는 것이 아닌 θ 의 함수 형태를 추정해야 하는 일이 자주 발생한다.

추정하고자 하는 파라미터가 $\alpha = g(\theta)$ 와 같이 θ 에 대한 함수이며 r 차원의 함수의 함수일 경우 다음 수식을 만족해야 한다. (자세한 유도 과정은 Appendix 3B 참조)

$$\mathbf{C}_{\hat{\alpha}} - \frac{\partial g(\theta)}{\partial \theta} \mathbf{I}^{-1}(\theta) \frac{\partial g(\theta)^\top}{\partial \theta} \geq 0 \quad (81)$$

- $(\cdot) \geq 0$: 좌항에 있는 행렬이 positive semidefinite임을 의미한다.

자코비안 행렬 $\frac{\partial g(\theta)}{\partial \theta}$ 는 $r \times p$ 크기의 행렬이며 다음과 같이 정의된다.

$$\frac{\partial g(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & \frac{\partial g_1(\theta)}{\partial \theta_2} & \dots & \frac{\partial g_1(\theta)}{\partial \theta_p} \\ \frac{\partial g_2(\theta)}{\partial \theta_1} & \frac{\partial g_2(\theta)}{\partial \theta_2} & \dots & \frac{\partial g_2(\theta)}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_r(\theta)}{\partial \theta_1} & \frac{\partial g_r(\theta)}{\partial \theta_2} & \dots & \frac{\partial g_r(\theta)}{\partial \theta_p} \end{bmatrix} \quad (82)$$

3.5.1 Example 3.8 - CRLB for Signal-to-Noise Ratio

다음과 같은 관측 데이터가 주어졌다고 하자.

$$x[n] = A + w[n] \quad \text{where } w[n] \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (83)$$

$\theta = [A, \sigma^2]^\top$ 는 미지의 파라미터이며 우리가 추정하고자 하는 값은 $\alpha = \frac{A^2}{\sigma^2}$ 라고 하자. 이러한 α 를 signal to noise ratio(SNR)이라고 한다. $\alpha = g(\theta) = \theta_1^2/\theta_2 = A^2/\sigma^2$ 과 같이 나타낼 수 있으므로 이를 통해 Fisher information 행렬을 표현하면 다음과 같다.

$$\mathbf{I}(\theta) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix} \quad (84)$$

자코비안 행렬 $\frac{\partial g(\theta)}{\partial \theta}$ 은 다음과 같다.

$$\begin{aligned} \frac{\partial g(\theta)}{\partial \theta} &= \begin{bmatrix} \frac{\partial g(\theta)}{\partial \theta_1} & \frac{\partial g(\theta)}{\partial \theta_2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial g(\theta)}{\partial A} & \frac{\partial g(\theta)}{\partial \sigma^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{2A}{\sigma^2} & -\frac{A^2}{\sigma^4} \end{bmatrix} \end{aligned} \quad (85)$$

따라서 (81) 우측 항은 다음과 같이 구할 수 있다.

$$\begin{aligned} \frac{\partial g(\theta)}{\partial \theta} \mathbf{I}^{-1}(\theta) \frac{\partial g(\theta)^\top}{\partial \theta} &= \begin{bmatrix} \frac{2A}{\sigma^2} & -\frac{A^2}{\sigma^4} \end{bmatrix} \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^2}{N} \end{bmatrix} \begin{bmatrix} \frac{2A}{\sigma^2} \\ -\frac{A^2}{\sigma^4} \end{bmatrix} \\ &= \frac{4A^2}{N\sigma^2} + \frac{2A^4}{N\sigma^4} \\ &= \frac{4\alpha + 2\alpha^2}{N} \end{aligned} \quad (86)$$

따라서 SNR 추정값 $\hat{\alpha}$ 에 대한 CRLB는 다음과 같다.

$$\text{var}(\hat{\alpha}) \geq \frac{4\alpha + 2\alpha^2}{N} \quad (87)$$

3.6 CRLB for the General Gaussian Case

이전 섹션에서 스칼라, 벡터 파라미터에 대한 CRLB를 알아보았다면 이번 섹션에서는 일반 가우시안 케이스에서 CRLB에 대해 알아본다. 일반 가우시안 케이스의 관측 데이터는 다음과 같이 나타낼 수 있다.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{C}(\boldsymbol{\theta})) \quad (88)$$

위와 같이 평균과 표준편차는 파라미터 $\boldsymbol{\theta}$ 에 대해 종속적이다. 일반 가우시안 케이스에서 Fisher information 행렬은 다음과 같이 나타낼 수 있다. (자세한 유도 과정은 Appendix 3C 참조)

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = \left[\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i} \right]^T \mathbf{C}^{-1}(\boldsymbol{\theta}) \left[\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_j} \right] + \frac{1}{2} \text{tr} \left[\mathbf{C}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_i} \mathbf{C}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_j} \right] \quad (89)$$

$$\text{where } \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i} = \begin{bmatrix} \frac{\partial [\boldsymbol{\mu}(\boldsymbol{\theta})]_1}{\partial \theta_i} \\ \frac{\partial [\boldsymbol{\mu}(\boldsymbol{\theta})]_2}{\partial \theta_i} \\ \vdots \\ \frac{\partial [\boldsymbol{\mu}(\boldsymbol{\theta})]_N}{\partial \theta_i} \end{bmatrix} \quad (90)$$

$$\text{and } \frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_i} = \begin{bmatrix} \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{11}}{\partial \theta_i} & \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{12}}{\partial \theta_i} & \dots & \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{1N}}{\partial \theta_i} \\ \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{21}}{\partial \theta_i} & \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{22}}{\partial \theta_i} & \dots & \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{2N}}{\partial \theta_i} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{N1}}{\partial \theta_i} & \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{N2}}{\partial \theta_i} & \dots & \frac{\partial [\mathbf{C}(\boldsymbol{\theta})]_{NN}}{\partial \theta_i} \end{bmatrix} \quad (91)$$

스칼라 파라미터의 경우라면 관측 데이터는 다음과 같이 표현할 수 있다.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}(\theta), \mathbf{C}(\theta)) \quad (92)$$

Fisher information 행렬은 다음과 같이 간단하게 표현할 수 있다.

$$\mathbf{I}(\theta) = \left[\frac{\partial \boldsymbol{\mu}(\theta)}{\partial \theta} \right]^T \mathbf{C}^{-1}(\theta) \left[\frac{\partial \boldsymbol{\mu}(\theta)}{\partial \theta} \right] + \frac{1}{2} \text{tr} \left[\mathbf{C}^{-1}(\theta) \frac{\partial \mathbf{C}(\theta)}{\partial \theta} \right] \quad (93)$$

3.6.1 Example 3.11 - Random DC Level in WGN

다음과 같은 관측 데이터가 주어졌다고 가정하자.

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1 \quad (94)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$
- $A \sim \mathcal{N}(0, \sigma_A^2)$

관측 데이터 $\mathbf{x} = [x[0], x[1], \dots, x[N-1]]^T$ 이 주어졌을 때 이는 평균이 0이며 다음과 같은 $N \times N$ 크기의 공분산 행렬을 가진다.

$$\begin{aligned} [\mathbf{C}(\sigma_A^2)]_{ij} &= \mathbb{E}[x[i-1]x[j-1]] \\ &= \mathbb{E}[(A + w[i-1])(A + w[j-1])] \\ &= \sigma_A^2 + \sigma^2 \delta_{ij} \end{aligned} \quad (95)$$

$$\therefore \mathbf{C}(\sigma_A^2) = \sigma_A^2 \mathbf{1}\mathbf{1}^T + \sigma^2 \mathbf{I} \quad (96)$$

- $\mathbf{1} = [1, 1, \dots, 1]^T$

Woodbury identity 공식에 의해 공분산의 역행렬은 다음과 같다. 자세한 내용은 해당 링크를 참조하면 된다.

$$\mathbf{C}^{-1}(\sigma_A^2) = \frac{1}{\sigma^2} \left(\mathbf{I} - \frac{\sigma_A^2}{\sigma^2 + N\sigma_A^2} \mathbf{1}\mathbf{1}^T \right) \quad (97)$$

공분산의 편미분은 다음과 같다.

$$\frac{\partial \mathbf{C}(\sigma_A^2)}{\partial \sigma_A^2} = \mathbf{1}\mathbf{1}^\top \quad (98)$$

따라서 둘을 합치면 다음 공식이 성립한다.

$$\mathbf{C}^{-1}(\sigma_A^2) \frac{\partial \mathbf{C}(\sigma_A^2)}{\partial \sigma_A^2} = \frac{1}{\sigma^2 + N\sigma_A^2} \mathbf{1}\mathbf{1}^\top \quad (99)$$

최종적으로 (93) 식은 다음과 같이 쓸 수 있다.

$$\begin{aligned} \mathbf{I}(\sigma_A^2) &= \frac{1}{2} \text{tr} \left[\left(\frac{1}{\sigma^2 + N\sigma_A^2} \right)^2 \mathbf{1}\mathbf{1}^\top \mathbf{1}\mathbf{1}^\top \right] \\ &= \frac{N}{2} \left(\frac{1}{\sigma^2 + N\sigma_A^2} \right)^2 \text{tr}(\mathbf{1}\mathbf{1}^\top) \\ &= \frac{1}{2} \left(\frac{N}{\sigma^2 + N\sigma_A^2} \right)^2 \end{aligned} \quad (100)$$

따라서 σ_A^2 에 대한 CRLB는 다음과 같이 나타낼 수 있다.

$$\text{var}(\sigma_A^2) \geq 2 \left(\sigma_A^2 + \frac{\sigma^2}{N} \right)^2 \quad (101)$$

위 식에서 보다시피 $N \rightarrow \infty$ 가 되어도 CRLB는 $2\sigma_A^4$ 이하로 내려가지 않는다. 이는 각각의 개별 관측 데이터가 전부 A 에 대한 정보를 포함하고 있기 때문에 A 의 분산 값이 항상 반영되기 때문이다.

4 Linear Models

일반적으로 MVUE를 찾는 과정은 쉬운 작업이 아니다. 하지만 다행이도 많은 신호 처리 문제들은 선형(linear) 데이터 모델의 형태를 가지는데 이러한 특성이 쉽게 MVUE를 결정할 수 있도록 도와준다. 선형 모델임이 확인되면 MVUE 값이 즉시 명백해지는 것 뿐만 아니라 통계적인 성능도 자연스럽게 따라온다. 그러므로 최적의 추정값을 찾는 열쇠는 문제를 선형 모델로 구조화하여 그 고유한 특성을 활용하는 것이 중요하다.

4.1 Definition and Properties

다음과 같은 line fitting 문제가 주어졌다고 하자.

$$x[n] = A + Bn + w[n] \quad n = 0, 1, \dots, N-1 \quad (102)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

이 때, y 절편의 값 A 와 직선의 기울기 B 의 값을 찾아야 한다. 위 식을 행렬 형태로 작성하면 다음과 같다.

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \quad (103)$$

각 원소를 펼쳐보면 다음과 같다.

$$\begin{aligned} \mathbf{x} &= [x[0], x[1], \dots, x[N-1]]^\top \\ \mathbf{w} &= [w[0], w[1], \dots, w[N-1]]^\top \\ \boldsymbol{\theta} &= [A, B]^\top \\ \mathbf{H} &= \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & N-1 \end{bmatrix} \end{aligned} \quad (104)$$

행렬 \mathbf{H} 는 $N \times 2$ 의 크기를 가지며 일반적으로 관측 함수(observation matrix)라고 부른다. 데이터 \mathbf{x} 는 파라미터 $\boldsymbol{\theta}$ 가 관측 행렬 \mathbf{H} 를 통과한 다음에 관측되기 때문이다. 노이즈 벡터도 $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ 와 같이 벡터 형태로 나타낼 수 있다. [이러한 \(103\)의 형태를 선형 모델\(liinear model\)이라고 부른다.](#) 선형모델의 노이즈는 일반적으로 가우시안 형태를 가진다.

앞서 3장의 Theorem 3.2에서 본 것처럼 $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$ 가 MVUE가 되려면 다음 식을 만족해야 한다.

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}) \quad (105)$$

그리고 위 식을 만족할 때 $\hat{\theta}$ 의 분산은 $\mathbf{I}^{-1}(\theta)$ 가 된다. 이러한 Theorem 3.2의 조건들을 선형 모델에 대해 확장하면 다음과 같다.

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \left[-\ln(2\pi\sigma^2)^{\frac{N}{2}} - \frac{1}{2\sigma^2} (\mathbf{x} - \mathbf{H}\theta)^\top (\mathbf{x} - \mathbf{H}\theta) \right] \\ &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \theta} [\mathbf{x}^\top \mathbf{x} - 2\mathbf{x}^\top \mathbf{H}\theta + \theta^\top \mathbf{H}^\top \mathbf{H}\theta]\end{aligned}\quad (106)$$

위 식에 다음과 같은 등식을 적용한다.

$$\begin{aligned}\frac{\partial \mathbf{b}^\top \theta}{\partial \theta} &= \mathbf{b} \\ \frac{\partial \theta^\top \mathbf{A} \theta}{\partial \theta} &= 2\mathbf{A}\theta\end{aligned}\quad (107)$$

- \mathbf{A} : 임의의 대칭(symmetric) 행렬

$$\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = \frac{1}{\sigma^2} [\mathbf{H}^\top \mathbf{x} - \mathbf{H}^\top \mathbf{H}\theta] \quad (108)$$

만약 $\mathbf{H}^\top \mathbf{H}$ 의 역행렬이 존재한다면 식은 다음과 같다.

$$\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = \frac{\mathbf{H}^\top \mathbf{H}}{\sigma^2} [(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x} - \theta] \quad (109)$$

위 식은 (105)의 형태와 정확히 일치한다.

$$\boxed{\begin{aligned}\hat{\theta} &= (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x} \\ \mathbf{I}(\theta) &= \frac{\mathbf{H}^\top \mathbf{H}}{\sigma^2}\end{aligned}} \quad (110)$$

따라서 MVUE를 만족하는 $\hat{\theta}$ 의 분산은 $\mathbf{I}^{-1}(\theta)$ 가 된다.

$$\boxed{\mathbf{C}_{\hat{\theta}} = \mathbf{I}^{-1}(\theta) = \sigma^2 (\mathbf{H}^\top \mathbf{H})^{-1}} \quad (111)$$

$$\boxed{\therefore \hat{\theta} \sim \mathcal{N}(\theta, \sigma^2 (\mathbf{H}^\top \mathbf{H})^{-1})} \quad (112)$$

4.1.1 Example 4.2 - Fourier Analysis

많은 신호는 주기적인 특성을 지닌다. 이러한 강한 주기적 특성은 푸리에 해석(Fourier analysis)를 통해 수학적으로 모델링할 수 있다. 푸리에 계수(coefficient)로 해당 주파수 성분이 강하게 포함되어 있다는 의미를 지닌다. 해당 예제에서는 푸리에 해석을 하는 과정이 선형 모델 파라미터를 추정하는 것과 동일한 과정임을 보인다. 다음과 같은 정현파 신호에 가우시안 노이즈가 추가된 예제를 보자.

$$x[n] = \sum_{k=1}^M a_k \cos\left(\frac{2\pi kn}{N}\right) + \sum_{k=1}^M b_k \sin\left(\frac{2\pi kn}{N}\right) + w[n] \quad n = 0, 1, \dots, N-1 \quad (113)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

주파수 성분은 $f_1 = 1/N$ 부터 $f_k = k/N$ 성분이 복합적으로 섞여 있는 것으로 가정한다. 여기서 우리는 a_k, b_k 계수를 추정해야 한다. 벡터 파라미터를 다음과 같이 정의한다.

$$\theta = [a_1, a_2, \dots, a_M, b_1, b_2, \dots, b_M]^\top \quad (114)$$

그리고 \mathbf{H} 행렬을 다음과 같이 정의한다.

$$\mathbf{H} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \cos\left(\frac{2\pi}{N}\right) & \cdots & \cos\left(\frac{2\pi M}{N}\right) & \sin\left(\frac{2\pi}{N}\right) & \cdots & \sin\left(\frac{2\pi M}{N}\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos\left(\frac{2\pi(N-1)}{N}\right) & \cdots & \cos\left(\frac{2\pi M(N-1)}{N}\right) & \sin\left(\frac{2\pi(N-1)}{N}\right) & \cdots & \sin\left(\frac{2\pi M(N-1)}{N}\right) \end{bmatrix} \quad (115)$$

위 식에서 행렬의 크기는 $N \times 2M$ 이며 $p = 2M$ 이 된다. \mathbf{H} 가 $N > p$ 를 만족하려면 $M < N/2$ 가 되어야 한다. \mathbf{H} 의 각 열벡터는 DFT 특성에 의해 서로 직교(orthogonal)하므로 계산을 편하게 하기 위해 이를 열벡터로 표시한다.

$$\mathbf{H} = [\mathbf{h}_1 \quad \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_{2M}] \quad (116)$$

직교성(orthogonality)에 의해 다음 식을 만족한다.

$$\mathbf{h}_i^T \mathbf{h}_j = 0 \quad \text{for } i \neq j \quad (117)$$

$\mathbf{H}^T \mathbf{H}$ 를 전개해보면 다음과 같다.

$$\begin{aligned} \mathbf{H}^T \mathbf{H} &= \begin{bmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_{2M}^T \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \cdots & \mathbf{h}_{2M} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{h}_1^T \mathbf{h}_1 & \mathbf{h}_1^T \mathbf{h}_2 & \cdots & \mathbf{h}_1^T \mathbf{h}_{2M} \\ \mathbf{h}_2^T \mathbf{h}_1 & \mathbf{h}_2^T \mathbf{h}_2 & \cdots & \mathbf{h}_2^T \mathbf{h}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{2M}^T \mathbf{h}_1 & \mathbf{h}_{2M}^T \mathbf{h}_2 & \cdots & \mathbf{h}_{2M}^T \mathbf{h}_{2M} \end{bmatrix} \end{aligned} \quad (118)$$

위 식은 직교성 특징으로 인해 대각 행렬(diagonal matrix)가 되고 매우 쉽게 역행렬을 구할 수 있다. DFT의 직교 성질은 다음과 같다.

$$\begin{aligned} \sum_{n=0}^{N-1} \cos\left(\frac{2\pi in}{N}\right) \cos\left(\frac{2\pi jn}{N}\right) &= \frac{N}{2} \delta_{ij} \\ \sum_{n=0}^{N-1} \sin\left(\frac{2\pi in}{N}\right) \sin\left(\frac{2\pi jn}{N}\right) &= \frac{N}{2} \delta_{ij} \\ \sum_{n=0}^{N-1} \cos\left(\frac{2\pi in}{N}\right) \sin\left(\frac{2\pi jn}{N}\right) &= 0 \quad \text{for all } i, j \end{aligned} \quad (119)$$

직교성 특징을 적용하면 다음과 같이 간단한 식으로 변한다.

$$\mathbf{H}^T \mathbf{H} = \begin{bmatrix} \frac{N}{2} & 0 & \cdots & 0 \\ 0 & \frac{N}{2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{N}{2} \end{bmatrix} = \frac{N}{2} \mathbf{I} \quad (120)$$

$\hat{\theta}$ 가 MVUE이면 다음 식을 만족해야 한다.

$$\begin{aligned} \hat{\theta} &= (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \\ &= \frac{2}{N} \mathbf{H}^T \mathbf{x} \\ &= \frac{2}{N} \begin{bmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_{2M}^T \end{bmatrix} \mathbf{x} \\ &= \begin{bmatrix} \frac{2}{N} \mathbf{h}_1^T \mathbf{x} \\ \vdots \\ \frac{2}{N} \mathbf{h}_{2M}^T \mathbf{x} \end{bmatrix} \end{aligned} \quad (121)$$

따라서 a_k, b_k 는 다음과 같이 추정할 수 있다.

$$\begin{aligned} \hat{a}_k &= \frac{2}{N} \sum_{n=0}^{N-1} x[n] \cos\left(\frac{2\pi kn}{N}\right) \\ \hat{b}_k &= \frac{2}{N} \sum_{n=0}^{N-1} x[n] \sin\left(\frac{2\pi kn}{N}\right) \end{aligned} \quad (122)$$

이는 DFT 계수와 동일하다. 추정값의 기대값과 분산을 구해보면 다음과 같다.

$$\begin{aligned}
\mathbb{E}(\hat{a}_k) &= a_k \\
\mathbb{E}(\hat{b}_k) &= b_k \\
\mathbf{C}_{\hat{\theta}} &= \sigma^2 (\mathbf{H}^\top \mathbf{H})^{-1} \\
&= \sigma^2 \left(\frac{N}{2} \mathbf{I} \right)^{-1} \\
&= \frac{2\sigma^2}{N} \mathbf{I}
\end{aligned} \tag{123}$$

추정값의 분산 $\mathbf{C}_{\hat{\theta}}$ 이 대각 행렬이므로 각 파라미터들은 서로 독립적이다. 이러한 독립적 특성으로 인해 \mathbf{H} 의 열 벡터가 서로 직교하여 $\mathbf{H}^\top \mathbf{H}$ 가 매우 단순하게 계산되었다. 만약 주파수 성분을 임의의 성분으로 변경하면 직교성이 상실되어 MVUE를 구하는 과정이 매우 복잡해진다.

4.2 Extension to the Linear Model

일반적인 경우 선형 모델의 노이즈는 WGN이 아니다. 선형 모델을 일반적인 경우로 확장하면 노이즈는 다음과 같이 모델링할 수 있다.

$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{C}) \tag{124}$$

\mathbf{C} 행렬은 더 이상 스칼라 값이 곱해진 대각 identity 행렬이 아니다. 노이즈는 0 이상의 값을 가지기 때문에 \mathbf{C} 는 positive definite 행렬이라고 가정할 수 있고 따라서 \mathbf{C}^{-1} 또한 positive definite이다. 이를 다음과 같이 분해할 수 있다.

$$\mathbf{C}^{-1} = \mathbf{D}^\top \mathbf{D} \tag{125}$$

\mathbf{D} 행렬은 $N \times N$ 크기의 역행렬이 존재하는 행렬이다. \mathbf{D} 행렬은 기존 노이즈를 whitening 변환해주는데 사용된다.

$$\begin{aligned}
\mathbb{E}[(\mathbf{D}\mathbf{w})(\mathbf{D}\mathbf{w})^\top] &= \mathbf{D}\mathbf{C}\mathbf{D}^\top \\
&= \mathbf{D}\mathbf{D}^{-1}\mathbf{D}^{-\top}\mathbf{D}^\top \\
&= \mathbf{I}
\end{aligned} \tag{126}$$

일반적인 선형 모델은 다음과 같다.

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{127}$$

여기에서 whitening을 위해 다음과 같은 변환을 거친다.

$$\begin{aligned}
\mathbf{x}' &= \mathbf{D}\mathbf{x} \\
&= \mathbf{D}\mathbf{H}\boldsymbol{\theta} + \mathbf{D}\mathbf{w} \\
&= \mathbf{H}'\boldsymbol{\theta} + \mathbf{w}'
\end{aligned} \tag{128}$$

위 식은 $\mathbf{w}' \sim \mathcal{N}(0, \mathbf{I})$ 와 같이 whitening 되었다. 다음 순서는 앞서 설명한 선형 모델 공식과 동일하다.

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= (\mathbf{H}'^\top \mathbf{H}')^{-1} \mathbf{H}'^\top \mathbf{x}' \\
&= (\mathbf{H}'^\top \mathbf{D}^\top \mathbf{D} \mathbf{H}')^{-1} \mathbf{H}'^\top \mathbf{D}^\top \mathbf{D} \mathbf{x}
\end{aligned} \tag{129}$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}'^\top \mathbf{C}^{-1} \mathbf{H}')^{-1} \mathbf{H}'^\top \mathbf{C}^{-1} \mathbf{x} \tag{130}$$

분산도 동일한 과정을 거친다.

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}'^\top \mathbf{H}')^{-1} \tag{131}$$

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}'^\top \mathbf{C}^{-1} \mathbf{H}')^{-1} \tag{132}$$

5 General Minimum Variance Unbiased Estimation

이전 섹션에서 CRLB를 통해 추정값이 efficient함을 알 수 있고 efficient한 추정값은 MVUE가 되는 것을 알 수 있었다. 그리고 선형 모델(linear model)을 사용하여 다양한 예제를 확인하였다. 하지만 만약 efficient한 추정값이 존재하지 않더라도 MVUE를 찾는 것에 관심이 있을 수 있다. 이번 섹션에서는 이러한 관심사를 확인할 수 있는 **Rao-Blackwell-Lehmann-Scheffe(RBLS)** 이론에 대해 배우고 이를 위한 충분통계량(sufficient statistics)의 개념에 대해 배운다. RBLS를 사용하면 많은 경우 단순히 pdf를 보는 것 만으로도 MVUE 인지 여부를 판단할 수 있게 된다.

5.1 Sufficient Statistics

이전 챕터에서 DC Level A의 추정 문제를 다시 보면 다음과 같다.

$$x[n] = A + w[n] \quad (133)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

A 의 추정값 \hat{A} 는 다음과 같이 데이터의 평균으로 구하였다.

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad (134)$$

\hat{A} 는 MVUE이므로 최소 분산 $\frac{\sigma^2}{N}$ 을 가진다. 만약 다음과 같은 추정값을 사용한다고 가정하자.

$$\check{A} = x[0] \quad (135)$$

\check{A} 는 편향되지 않았다는 것(unbiased)을 알 수 있지만 \check{A} 의 분산은 σ^2 로 MVUE의 분산 $\frac{\sigma^2}{N}$ 대비 크다는 것을 알 수 있다. 직관적으로 다른 데이터 $x[1], x[2], \dots, x[N-1]$ 을 사용하지 않기 때문에 A 에 대한 정보를 상실하여 좋은 추정값이 아님을 알 수 있다. 여기서 어떤 데이터가 A 의 정보를 많이 포함하고 있을까? 또는 어떤 데이터가 A 를 추정하는데 충분할까? 아래와 같이 추정에 사용 가능한 집합들이 주어졌다고 가정해보자.

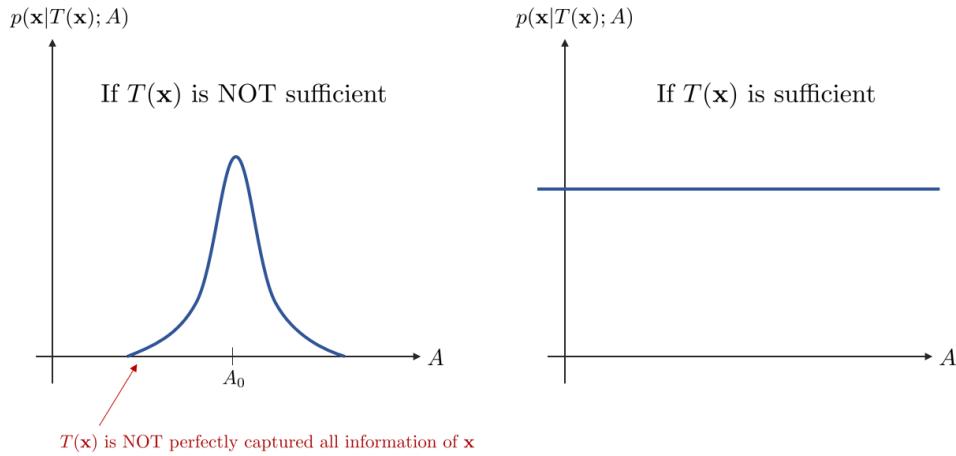
$$\begin{aligned} S_1 &= \{x[0], x[1], \dots, x[N-1]\} \\ S_2 &= \{x[0] + x[1], x[2], x[3], \dots, x[N-1]\} \\ S_3 &= \left\{ \sum_{n=0}^{N-1} x[n] \right\} \end{aligned} \quad (136)$$

S_1 은 기존 데이터 집합으로써 항상 충분한 데이터를 가지고 있다. S_2 와 S_3 또한 충분한 데이터를 가지고 있다. 이외에도 예제를 만족하는 수 많은 충분한 데이터 집합들이 있지만 우리는 이 중 가장 최소한의 크기를 지닌 집합을 원한다. S_1, S_2, S_3 모두 통계적 관점에서 sufficient하다고 볼 수 있지만 이 중 가장 작은 크기의 집합인 S_3 는 특별히 **minimal sufficient statistics(또는 충분통계량)**라고 한다. A 를 추정할 때 $\sum_{n=0}^{N-1} x[n]$ 은 각 데이터에 대한 모든 정보를 포함하고 있으므로 개별 데이터를 알지 못해도 추정이 가능하다.

충분통계량을 자세히 알기 위해 pdf를 보면 다음과 같다.

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \quad (137)$$

위 식에 임의의 통계량을 $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$ 이 주어진 경우 pdf $p(\mathbf{x}|T(\mathbf{x}); A)$ 가 파라미터 A 에 더 이상 종속적이지 않다면 $T(\mathbf{x})$ 는 파라미터를 추정하는데 충분한 통계량(i.e., 충분통계량)이라고 정의한다.



5.1.1 Example 5.1 - Verification of a Sufficient Statistic

(133)를 사용하여 충분통계량이 A 에 독립적임을 증명해보자. 통계량 $T(\mathbf{x}) = T_0$ 이 주어진 경우 조건부 pdf는 다음과 같이 전개할 수 있다.

$$p(\mathbf{x}|T(\mathbf{x}) = T_0; A) = \frac{p(\mathbf{x}, T(\mathbf{x}) = T_0; A)}{p(T(\mathbf{x}) = T_0; A)} \quad (138)$$

- $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$

위 식의 우측 상단의 joint pdf는 다음과 같으므로 Dirac delta function을 사용하여 분리할 수 있다. (자세한 내용은 Appendix 5A 참조)

$$p(\mathbf{x}|T(\mathbf{x}) = T_0; A) = \frac{p(\mathbf{x}; A)\delta(T(\mathbf{x}) - T_0)}{p(T(\mathbf{x}) = T_0; A)} \quad (139)$$

통계량은 정의 상 $T(\mathbf{x}) \sim \mathcal{N}(NA, N\sigma^2)$ 을 만족하므로 위 식은 다음과 같이 전개할 수 있다.

$$\begin{aligned} p(\mathbf{x}; A)\delta(T(\mathbf{x}) - T_0) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \delta(T(\mathbf{x}) - T_0) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \left(\sum_{n=0}^{N-1} x^2[n] - 2AT(\mathbf{x}) + NA^2 \right) \right] \delta(T(\mathbf{x}) - T_0) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \left(\sum_{n=0}^{N-1} x^2[n] - 2AT_0 + NA^2 \right) \right] \delta(T(\mathbf{x}) - T_0) \end{aligned} \quad (140)$$

(140)을 (139)에 대입하면 다음과 같다.

$$\begin{aligned} p(\mathbf{x}|T(\mathbf{x}) = T_0; A) &= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n] \right] \exp \left[-\frac{1}{2\sigma^2} (-2AT_0 + NA^2) \right]}{\sqrt{2\pi N\sigma^2} \exp \left[-\frac{1}{2N\sigma^2} (T_0 - NA)^2 \right]} \delta(T(\mathbf{x}) - T_0) \\ &= \frac{\sqrt{N}}{(2\pi\sigma^2)^{\frac{N-1}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n] \right] \exp \left[\frac{T_0^2}{2N\sigma^2} \right] \delta(T(\mathbf{x}) - T_0) \end{aligned} \quad (141)$$

위 식 마지막 줄에서 보다시피 파라미터 A 가 존재하지 않으므로 더 이상 A 에 종속적이지 않다. 따라서 $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$ 은 A 를 추정하는데 충분한 통계량(i.e., 충분통계량)인 것을 알 수 있다.

5.2 Finding Sufficient Statistics

5.2.1 Theorem 5.1 (Neyman-Fisher Factorization)

임의의 통계량 $T(\mathbf{x})$ 이 파라미터 θ 에 대한 충분통계량이면 $p(\mathbf{x}; \theta)$ 는 반드시 다음과 같은 형태로 분해(factorization) 할 수 있다. 해당 정리의 유도 과정은 Appendix 5A를 참조하면 된다.

$$p(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x}) \quad (142)$$

- $g(\cdot)$: $T(\mathbf{x})$ 를 포함하는 \mathbf{x} 에 종속적인 임의의 함수
- $h(\cdot)$: \mathbf{x} 에만 종속적인 임의의 함수

5.2.2 Example 5.2 - DC Level in WGN

앞서 DC Level A 의 추정 문제를 다시 보면 다음과 같다. 여기서 σ^2 는 이미 알고 있다고 가정한다.

$$x[n] = A + w[n] \quad (143)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

pdf의 exponential 항을 전개하면 다음과 같다.

$$\sum_{n=0}^{N-1} (x[n] - A)^2 = \sum_{n=0}^{N-1} x^2[n] - 2A \sum_{n=0}^{N-1} x[n] + NA^2 \quad (144)$$

따라서 $p(\mathbf{x}; \theta)$ 는 다음과 같이 분해할 수 있다.

$$p(\mathbf{x}; \theta) = \underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \left(NA^2 - 2A \sum_{n=0}^{N-1} x[n] \right) \right]}_{g(T(\mathbf{x}), A)} \underbrace{\exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n] \right]}_{h(\mathbf{x})} \quad (145)$$

위 식은 Neyman-Fisher Factorization 정리와 동일하게 분해가 되므로 따라서 $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$ 은 충분통계량이 된다. 만약 $T'(\mathbf{x}) = 2 \sum_{n=0}^{N-1} x[n]$ 통계량이 있다면 이를 위 pdf 식에 넣어도 똑같이 분해가 된다. 즉, $T'(\mathbf{x})$ 도 충분통계량이 된다. **따라서 $T(\mathbf{x})$ 와 일대일 함수(one-to-one function) 관계에 있는 모든 함수들은 충분통계량이 된다. 이는 오직 $T(\mathbf{x})$ 와 일대일 함수 변환 관계에서만 성립한다.**

5.2.3 Proof of the Neymann-Fisher Factorization

TBD

5.3 Using Sufficiency to Find the MVU Estimator

5.3.1 Example 5.5 - DC Level in WGN

앞서 DC Level A 의 추정 문제로 돌아가보자. 여기서 σ^2 는 이미 알고 있다고 가정한다.

$$x[n] = A + w[n] \quad (146)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

우리는 이미 $\hat{A} = \bar{x}$ 가 CRLB를 만족하는 efficient한 추정값이면서 동시에 MVUE인것을 알지만 이번 섹션에서는 RBLS 이론을 사용하여 이를 다시 구해보자 한다. **RBLS 이론을 사용하면 CRLB를 만족하지 않아서 추정값이 efficient하지 않은 경우에도 MVUE를 찾을 수 있다.** MVUE를 찾기 위해 이번 예제에서는 두 가지 다른 방법을 사용한다. 두 방법은 전개 과정은 다르지만 둘 다 충분통계량 $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$ 을 사용한다는 공통점이 있다.

Approach 1

임의의 불편추정값 $\check{A} = x[0]$ 가 주어졌다고 가정하자. 찾고자하는 추정값을 $\hat{A} = \mathbb{E}(\check{A}|T)$ 과 같이 정의한다. \hat{A} 의 기대값은 $p(\hat{A}|T)$ 를 사용하여 구할 수 있다. \hat{A} 를 자세히 나타내면 다음과 같다.

$$\hat{A} = \mathbb{E}(x[0] | \sum_{n=0}^{N-1} x[n]) \quad (147)$$

이를 전개하기 위해서는 조건부(conditional) pdf를 사용해야 한다.

조건부 pdf에서 $x = x[0]$ 이고 $y = \sum_{n=0}^{N-1} x[n]$ 이다. 이를 자세히 전개하면 다음과 같다.

Tip

두 개의 확률변수 $[x, y]^\top$ 이 주어졌을 때 둘의 결합(joint) pdf의 평균은 $\mu = \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(y) \end{bmatrix}$ 이고 분산은 $\mathbf{C} = \begin{bmatrix} \text{var}(x) & \text{cov}(x, y) \\ \text{cov}(y, x) & \text{var}(y) \end{bmatrix}$ 이다. 두 변수의 조건부 pdf는 다음과 같이 유도할 수 있다.

$$\begin{aligned}\mathbb{E}(x|y) &= \int_{-\infty}^{\infty} xp(x|y)dx \\ &= \int_{-\infty}^{\infty} x \frac{p(x,y)}{p(y)} dx \\ &= \mathbb{E}(x) + \frac{\text{cov}(x,y)}{\text{var}(y)}(y - \mathbb{E}(y))\end{aligned}\quad (148)$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x[0] \\ \sum_{n=0}^{N-1} x[n] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}}_{\mathbf{L}} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} \quad (149)$$

$[x, y]^\top$ 은 가우시안 분포 $\mathcal{N}(\mu, \mathbf{C})$ 를 따른다고 할 때 μ, \mathbf{C} 는 각각 가우시안 벡터의 선형 변환(linear transformation) 형태이다.

$$\begin{aligned}\mu &= \mathbf{L}\mathbb{E}(\mathbf{x}) = \mathbf{L}A\mathbf{1} = \begin{bmatrix} A \\ NA \end{bmatrix} \\ \mathbf{C} &= \sigma^2 \mathbf{L} \mathbf{L}^\top = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & N \end{bmatrix}\end{aligned}\quad (150)$$

따라서 \hat{A} 는 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\hat{A} &= \mathbb{E}(x|y) \\ &= A + \frac{\sigma^2}{N\sigma^2} \left(\sum_{n=0}^{N-1} x[n] - NA \right) \\ &= \frac{1}{N} \sum_{n=0}^{N-1} x[n]\end{aligned}\quad (151)$$

따라서 \hat{A} 는 MVUE가 된다. 이는 수학적으로 조건부 확률의 기대값을 사용하기 때문에 상대적으로 다루기 어렵다.

Approach 2

MVUE를 구하기 위해 충분통계량 T 와 관련된 임의의 함수 $g(T)$ 를 찾는다. 즉 찾고자하는 추정값을 $\hat{A} = g(T)$ 와 같이 세팅한다. 이를 자세히 나타내면 다음과 같다.

$$\hat{A} = g\left(\sum_{n=0}^{N-1} x[n]\right) \quad (152)$$

위 함수를 자세히 보면 $g(x) = x/N$ 으로 세팅하는 것이 \hat{A} 를 MVUE로 만들 수 있다는 것을 쉽게 알 수 있다.

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad (153)$$

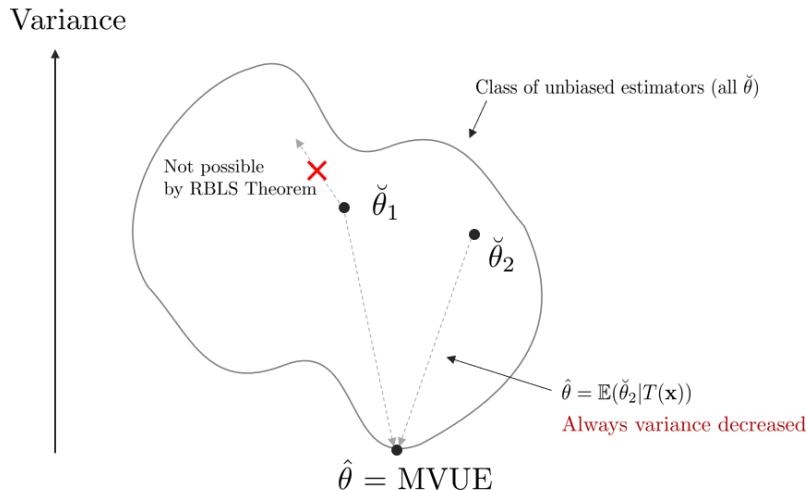
Approach 1,2 중 1이 수학적으로는 더 염밀하고 정확하지만 2를 사용하는 것이 쉽고 빠르게 MVUE를 찾을 수 있다는 것을 알 수 있다. 실제 예제에서는 2와 같은 방법을 많이 사용한다.

5.3.2 Theorem 5.2 (Rao-Blackwell-Lehmann-Scheffe)

$\check{\theta}$ 가 파라미터 θ 에 대한 불편추정값(unbiased estimator)이면서 $T(\mathbf{x})$ 가 θ 에 대한 충분통계량일 때, $\hat{\theta} = \mathbb{E}(\check{\theta}|T(\mathbf{x}))$ 는 다음과 같은 성질을 지닌다.

파라미터 θ 에 대한 유효한 추정값이면서 동시에 θ 에 종속적이지 않다. 편향되지 않았다. (unbiased) 모든 θ 에 대하여 $\check{\theta}$ 의 분산보다 $\hat{\theta}$ 의 분산이 작거나 같다. 그리고 만약 충분통계량이 complete한 경우, $\hat{\theta}$ 는 MVUE가 된다. 이에 대한 자세한 유도 과정은 Appendix 5B를 참조하면 된다.

이전 예제에서 우리는 $\hat{A} = \mathbb{E}(x[0]|\sum_{n=0}^{N-1} x[n]) = \bar{x}$ 가 A 에 대해 종속적이지도 않으면서 편향되지도 않고 분산도 $x[0]$ 보다 작은 것을 확인하였다. 따라서 RBLS 정리에 따라 충분통계량 $\sum_{n=0}^{N-1} x[n]$ 은 **complete한 충분통계량**이라고 정의할 수 있다. 충분통계량의 completeness에 대한 설명은 뒤에서 더 자세히 다룬다.



위 그림과 같이 파라미터 θ 를 추정하는 모든 불편추정값들의 집합이 있다고 가정하자. 이 때, $\mathbb{E}(\check{\theta}|T(\mathbf{x}))$ 값은 $\check{\theta}$ 보다 항상 작거나 같은 분산을 가진다. 조건부 기대값을 자세히 전개해보면 이는 $T(\mathbf{x})$ 에 대한 단일 함수라는 것을 알 수 있다.

$$\begin{aligned}\hat{\theta} &= \mathbb{E}(\check{\theta}|T(\mathbf{x})) \\ &= \int \check{\theta} p(\check{\theta}|T(\mathbf{x})) d\check{\theta} \\ &= g(T(\mathbf{x}))\end{aligned}\tag{154}$$

이전 섹션에서 충분통계량을 설명할 때 $p(\mathbf{x}|T(\mathbf{x}); \theta)$ 는 θ 에 독립적임을 설명하였다. 따라서 $\hat{\theta}$ 는 파라미터 θ 에 종속적이지 않다. (154)의 세 번째 줄이 Approach 1과 같은 조건부 pdf를 사용하지 않고 Approach 2와 같은 간단한 방법을 사용하는 이유이다.

$T(\mathbf{x})$ 가 **complete**하려면 $g(T(\mathbf{x}))$ 는 유일한 함수여야 한다. 다시 말하면 $g(T(\mathbf{x}))$ 로 구할 수 있는 추정값이 여러 개가 아닌 오직 $\hat{\theta}$ 는 하나만 존재해야 한다(**must be unique**). 앞선 예제의 Approach 2에서는 $g(\sum_{n=0}^{N-1} x[n]) = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$)이 유일한 함수인 경우이며 이 때 $\sum_{n=0}^{N-1} x[n]$ 는 **complete한 충분통계량**이라고 한다.

모든 불편추정값 $\check{\theta}$ 는 하나의 $\hat{\theta}$ 로 매핑될 수 있다. $\hat{\theta}$ 는 모든 불편추정값 중 분산이 가장 작은 추정값이 되므로 이는 곧 MVUE임을 의미한다. 따라서 어떤 불편추정값이든 RBLS를 적용하면 MVUE를 찾을 수 있다.

5.3.3 Example 5.6 - Completeness of a Sufficient Statistic

예제 5.5에서 A 를 추정할 때 충분통계량 $\sum_{n=0}^{N-1} x[n]$ 은 **complete한 충분통계량**으로써 오직 $\mathbb{E}[g(\sum_{n=0}^{N-1} x[n])] = A$ 와 같이 유일한 함수만 존재함을 알았다. 만약 $\mathbb{E}[h(\sum_{n=0}^{N-1} x[n])] = A$ 를 만족하는 함수 h 가 존재한다고 가정해보자. 그러면 다음 공식이 성립해야 한다.

$$\mathbb{E}[g(T) - h(T)] = A - A = 0 \quad \text{for all } A\tag{155}$$

충분통계량은 가우시안 분포 $T \sim \mathcal{N}(NA, N\sigma^2)$ 를 따르기 때문에 다음과 같이 전개할 수 있다.

$$\int_{-\infty}^{\infty} v(T) \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{1}{2N\sigma^2}(T-NA)^2\right] dT = 0 \quad \text{for all } A\tag{156}$$

$$- v(T) = g(T) - h(T)$$

위 식에서 $\tau = T/N$ 으로 치환하고 $v'(\tau) = v(N\tau)$ 로 치환하면 다음과 같다.

$$\int_{-\infty}^{\infty} v'(\tau) \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{N}{2\sigma^2}(A - \tau)^2\right] d\tau = 0 \quad \text{for all } A \quad (157)$$

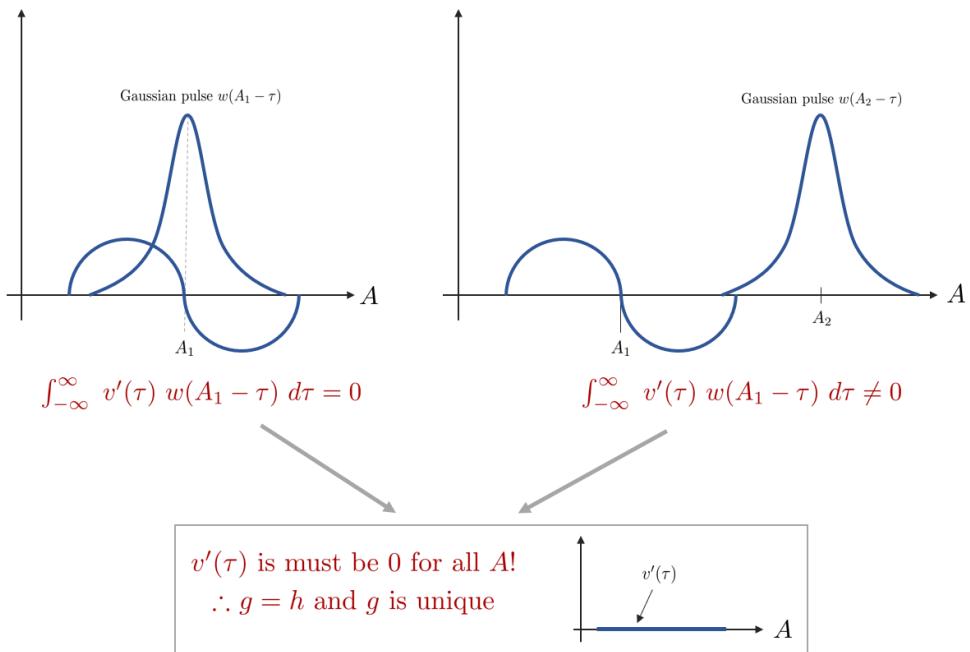
$$\boxed{\int_{-\infty}^{\infty} v'(\tau) w(A - \tau) d\tau \quad \text{for all } A} \quad (158)$$

위 식은 함수 $v'(\tau)$ 가 가우시안 펄스(pulse) $w(\tau)$ 와 컨볼루션 연산을 하는 것과 동일하다. 모든 A 에 대하여 0인 값을 $v'(\tau)$ 에 대해서도 동일하게 0의 값을 가진다. 시간 도메인에서 신호의 값이 0인 경우 컨볼루션의 푸리에 변환도 동일하게 0의 값을 가진다.

$$V'(f)W(f) = 0 \quad \text{for all } A \quad (159)$$

- $V'(f) = \mathcal{F}\{v'(\tau)\}$: 함수 $v'(\tau)$ 를 푸리에 변환한 값
- $W(f) = \mathcal{F}\{w(\tau)\}$: 함수 $w(\tau)$ 를 푸리에 변환한 값

가우시안 펄스 $w(\tau)$ 의 푸리에 변환 $W(f)$ 도 가우시안 펄스이므로 모든 주파수 f 에 대하여 0이 아닌 값을 가진다. 따라서 (159)를 만족하려면 $V'(\tau)$ 가 반드시 모든 주파수 f 에 대하여 0의 값을 가져야 한다. 따라서 $v'(\tau)$ 도 모든 τ 에 대하여 0의 값을 가져야 한다. 이는 $g = h$ 를 의미하며 따라서 함수 g 는 유일하다는 것을 알 수 있다.



5.3.4 Example 5.7 - Incomplete Sufficient Statistic

다음과 같은 단일 관측 데이터가 주어졌다고 가정하자

$$x[0] = A + w[0] \quad (160)$$

여기서 노이즈는 $w[0] \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ 와 같은 균일 분포를 가진다. 충분통계량은 $x[0]$ 가 되고 $x[0]$ 는 곧 A 의 불편추정값이 된다. 우리는 $g(x[0]) = x[0]$ 가 MVUE의 후보 중 하나라는 것을 알고 있다. 여기서 궁금한 것은 MVUE를 만족하는 것을 통해 충분통계량이 completeness한지를 알 수 있느냐는 것이다.

이전 예제와 동일하게 임의의 함수 $h(x[0]) = A$ 가 있다고 가정하자. 우리는 $h = g$ 를 만족하는지 여부를 봐야 한다. 만약 $h = g$ 라면 이전 예제와 동일하게 $x[0]$ 는 complete한 충분통계량이 된다. $v(T) = g(T) - h(T)$ 라고 하면 다음 공식이 성립한다.

$$\int_{-\infty}^{\infty} v(T) p(\mathbf{x}; A) d\mathbf{x} = 0 \quad \text{for all } A \quad (161)$$

위 예제에서는 $\mathbf{x} = x[0] = T$ 이므로 위 식은 아래와 같이 쓸 수 있다.

$$\int_{-\infty}^{\infty} v(T) p(T; A) dT = 0 \quad \text{for all } A \quad (162)$$

노이즈가 $w[0] \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ 인 균일 분포이므로 $p(T; A)$ 는 다음과 같다.

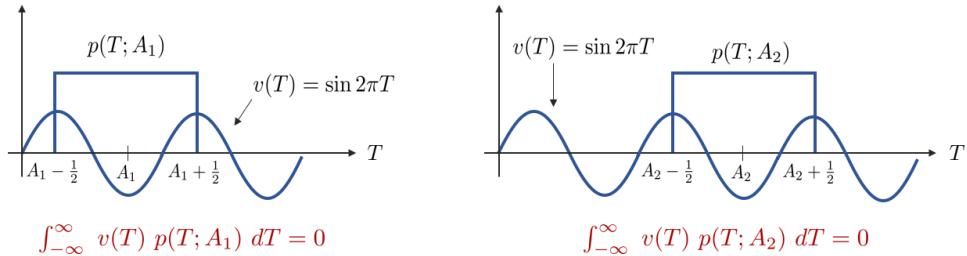
$$p(T; A) = \begin{cases} 1 & A - \frac{1}{2} \leq T \leq A + \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad (163)$$

따라서 위 식은 다음과 같이 쓸 수 있다.

$$\int_{A - \frac{1}{2}}^{A + \frac{1}{2}} v(T) dT = 0 \quad (164)$$

0이 아닌 주기함수 $v(T) = \sin 2\pi T$ 는 아래 그림과 같이 (164)을 만족한다.

$$v(T) = g(T) - h(T) = \sin 2\pi T \quad (165)$$



$v(T) = \sin 2\pi T$ satisfied for all A !

$\therefore g \neq h \rightarrow x[0]$ is not a complete sufficient statistics!

따라서 $h(T) = T - \sin 2\pi T$ 가 되고 불편추정값 $\hat{A} = x[0] - \sin 2\pi x[0]$ 을 얻을 수 있다. 결론적으로 충분통계량에 대한 함수 $g(T)$ 가 유일하지 않다면 불편추정값은 존재할 수 있으나 이는 complete하지 않다. Complete하지 않은 충분통계량은 RBLS 정리를 만족하지 않으므로 \hat{A} 는 반드시 MVUE임을 보장하지 않는다.

충분통계량의 completeness를 정리하면 다음과 같다. 아래와 같은 공식이 주어졌을 때

$$\boxed{\int_{-\infty}^{\infty} v(T)p(\mathbf{x}; A)d\mathbf{x} = 0 \quad \text{for all } A} \quad (166)$$

위 식이 모든 T 에 대하여 항상 $v(T) = 0$ 을 만족해야 충분통계량이 complete하다고 할 수 있다. 지금까지 배운 RBLS를 통하여 MVUE를 찾는 과정을 정리하면 다음과 같다.

1. 파라미터 θ 에 대해 유일한 충분통계량 $T(\mathbf{x})$ 를 Neymann-Fisher Factorization을 통해 찾는다.
2. 충분통계량이 complete한지 검사한다. 만약 그렇다면 RBLS를 적용할 수 있고 그렇지 않다면 적용할 수 없다.
3. $\hat{\theta} = g(T(\mathbf{x}))$ 를 만족하는 함수 g 를 찾는다. 만약 찾는다면 $\hat{\theta}$ 가 MVUE가 된다.
4. 3 과정 대신 $\hat{\theta} = \mathbb{E}(\check{\theta}|T(\mathbf{x}))$ 를 사용하여 MVUE를 구할 수도 있다.

일반적인 추정 문제에서 네 번째 과정의 조건부 확률을 구하는 것은 매우 어려운 작업이다. 이를 함수 형태로 표현하면 다음과 같다.

- RBLS{

 1. Do Neymann-Fisher factorization $p(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$
 2. Determine if $T(\mathbf{x})$ is complete.
 3. Find unique function $g(T(\mathbf{x}))$. \leftarrow Simple Ver.
 4. Find $\hat{\theta} = \mathbb{E}(\check{\theta}|T(\mathbf{x}))$. \leftarrow Complicated Ver.

}
- (167)

5.4 Extension to a Vector Parameter

지금까지 배운 내용들은 모두 스칼라 파라미터 θ 를 추정하는 문제였다. 이번 섹션에서는 이를 $p \times 1$ 크기의 벡터 파라미터 θ 로 확장하여 설명한다.

5.4.1 Example 5.10 and Example 5.11 - DC Level in WGN with Unknown Noise Power

다음과 같은 관측 데이터가 주어졌다고 가정하자.

$$x[n] = A + w[n] \quad (168)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

위 문제에서 우리가 추정하고자 하는 파라미터가 벡터 파라미터 $\theta = [A, \sigma^2]^T$ 인 경우를 생각해보자. Example 5.2와 5.3에서 구했다시피 충분통계량은 다음과 같다.

$$\mathbf{T}(\mathbf{x}) = \begin{bmatrix} T_1(\mathbf{x}) \\ T_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \sum_{n=0}^{N-1} x[n] \\ \sum_{n=0}^{N-1} x^2[n] \end{bmatrix} \quad (169)$$

충분통계량의 기대값을 구해보면 다음과 같다.

$$\begin{aligned} \mathbb{E}(\mathbf{T}(\mathbf{x})) &= \begin{bmatrix} NA \\ N\mathbb{E}(x^2[n]) \end{bmatrix} \\ &= \begin{bmatrix} NA \\ N(\sigma^2 + A^2) \end{bmatrix} \quad \text{← Second row is biased} \end{aligned} \quad (170)$$

위 행렬에서 첫 번째 행의 값은 bias를 제거하기 위해 $1/N$ 을 곱해주면 된다고 쉽게 예측할 수 있나 두 번째 행은 bias는 쉽게 제거할 수 없다. 두 번째 행의 bias를 제거하기 위해 $\mathbf{g}(\mathbf{T}(\mathbf{x}))$ 를 다음과 같이 선언한다.

$$\begin{aligned} \mathbf{g}(\mathbf{T}(\mathbf{x})) &= \begin{bmatrix} \frac{1}{N} T_1(\mathbf{x}) \\ \frac{1}{N} T_2(\mathbf{x}) - [\frac{1}{N} T_1(\mathbf{x})]^2 \end{bmatrix} \\ &= \begin{bmatrix} \bar{x} \\ \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] - \bar{x}^2 \end{bmatrix} \end{aligned} \quad (171)$$

$\mathbb{E}(\bar{x}) = A$ 이고

$$\mathbb{E}\left(\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] - \bar{x}^2\right) = \sigma^2 + A^2 - \mathbb{E}(\bar{x}^2) \quad (172)$$

이 된다. $\bar{x} \sim \mathcal{N}(A, \sigma^2/N)$ 이므로 $\mathbb{E}(\bar{x}^2)$ 는 다음과 같다.

$$\mathbb{E}(\bar{x}^2) = A^2 + \sigma^2/N \quad (173)$$

따라서 (172)를 다시 쓰면 다음과 같다.

$$\mathbb{E}\left(\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] - \bar{x}^2\right) = \sigma^2\left(1 - \frac{1}{N}\right) = \frac{N-1}{N}\sigma^2 \quad (174)$$

(174) 식에 $N/(N-1)$ 을 곱하면 bias가 사라지므로 따라서 (171)는 다음과 같다.

$$\begin{aligned} \mathbf{g}(\mathbf{T}(\mathbf{x})) &= \begin{bmatrix} \frac{1}{N} T_1(\mathbf{x}) \\ \frac{1}{N-1} T_2(\mathbf{x}) - N[\frac{1}{N} T_1(\mathbf{x})]^2 \end{bmatrix} \\ &= \begin{bmatrix} \bar{x} \\ \frac{1}{N-1} \sum_{n=0}^{N-1} x^2[n] - N\bar{x}^2 \end{bmatrix} \end{aligned} \quad (175)$$

하지만

$$\begin{aligned} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 &= \sum_{n=0}^{N-1} x^2[n] - 2 \sum_{n=0}^{N-1} x[n]\bar{x} + N\bar{x}^2 \\ &= \sum_{n=0}^{N-1} x^2[n] - N\bar{x}^2 \end{aligned} \quad (176)$$

위 등식을 사용하여 (175)을 다시 표현하면 최종적으로 MVUE $\hat{\theta}$ 를 얻을 수 있다.

$$\hat{\theta} = \left[\frac{1}{N-1} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 \right] \quad (177)$$

위와 같이 RBLS를 사용하면 MVUE $\hat{\theta}$ 를 얻을 수 있지만 g 함수에서 분산에 $1/(N-1)$ 를 곱해주어 자유도를 한 개 잃는다. 따라서 $\hat{\theta}$ 는 efficient하지 않다. [Hoel, Port, and Stone 1971] 문서를 참고하여 $\mathbf{C}_{\hat{\theta}}$ 를 구해보면 다음과 같다.

$$\mathbf{C}_{\hat{\theta}} = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N-1} \end{bmatrix} \quad (178)$$

위 식은 CRLB의 분산 $\mathbf{I}^{-1}(\hat{\theta})$ 보다 크다.

$$\mathbf{I}^{-1}(\hat{\theta}) = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix} \quad (179)$$

$$\mathbf{C}_{\hat{\theta}} > \mathbf{I}^{-1}(\hat{\theta}) \quad (180)$$

그리므로 CRLB를 사용하면 MVUE $\hat{\theta}$ 를 구할 수 없으나 RBLS를 사용하면 efficient하지 않은 MVUE $\hat{\theta}$ 를 얻을 수 있다.

6 Best Linear Unbiased Estimation

실제 추정 문제에서는 종종 MVUE가 존재한다고 해도 이를 찾을 수 없는 경우가 많다. 또한 데이터의 pdf를 알 수 조차 없는 경우가 대부분이다. 이런 경우에는 이전 챕터에서 배웠던 CRLB와 충분통계량을 적용할 수 없다. 최적의 MVUE를 찾을 수 없는 경우에 대비하여 suboptimal 추정값이라도 찾아야 한다. 하지만 suboptimal 추정값은 우리가 얼마나 추정 정확도에 대한 손해를 보고 있는지 알 수 없는 단점이 존재한다. 만약 suboptimal 추정값의 분산을 확정할 수 있고 그것이 우리 시스템의 사양을 충족한다면 이를 사용하는 현재 문제에 적절하다고 정당화 할 수 있다. **이런 경우 일반적인 접근법은 추정값을 데이터에 선형이며 최소 분산을 가지는 불편추정값 best linear unbiased estimator(BLUE)를 사용하는 것이다.** BLUE는 pdf의 첫번째와 두번째 모멘트만 사용하여 결정될 수 있다. 따라서 pdf를 정확히 모르는 경우에도 BLUE를 사용할 수 있기 때문에 실제 구현에 더 적합한 경우가 많다.

6.1 Definition of the BLUE

미지의 파라미터 θ 의 관측 데이터 $\{x[0], x[1], \dots, x[N-1]\}$ 이 주어졌고 이에 대한 pdf를 $p(\mathbf{x}; \theta)$ 라고 할 때 BLUE는 데이터의 선형 결합으로 표현할 수 있다.

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] \quad (181)$$

- a_n : 아직 결정되지 않은 미지의 상수

a_n 값이 어떻게 결정되느냐에 따라 다양한 추정값들이 존재할 수 있지만 BLUE는 이 중에서 가장 분산이 작은 불편추정값을 의미한다. a_n 을 결정하기 전에 최적성에 대한 분석이 필요하다. 우리가 찾고자 하는 추정값을 선형 추정값으로만 제한한다면 MVUE 또한 선형일 것이기 때문에 BLUE는 최적의 추정값이 된다. 예를 들어 WGN에서 DC Level의 값을 추정하는 Example 3.3과 같은 문제에서 MVUE는 다음과 같이 데이터의 평균을 의미한다.

$$\hat{\theta} = \bar{x} = \sum_{n=0}^{N-1} x[n] \quad (182)$$

이는 보다시피 데이터의 선형 결합으로 표현되어 있다. 따라서 우리가 선형 추정값에 대해서만 관심이 있는 경우 BLUE는 아무런 성능의 손실이 없는 MVUE가 될 수 있다. 반면에 Example 5.8과 같이 균일 분포 노이즈의 평균을 추정하고자 하는 경우 MVUE는 다음과 같다.

$$\hat{\theta} = \frac{N+1}{2N} \max x[n] \quad (183)$$

이는 데이터에 대한 비선형(nonlinear in the data) 추정값이다. 이러한 예제에서도 BLUE는 찾을 수 있지만 ($=\bar{x}$) 이는 suboptimal이 된다. 아쉽게도 pdf에 대한 정보가 없다면 suboptimal BLUE가 어느 정도 성능 손실을 보는지에 대해서는 알 수 없다.

마지막으로 BLUE를 사용하는 것이 까다로운 추정 문제도 있다. WGN의 파워를 추정하는 Example 3.6의 예제는 다음과 같은 MVUE를 가진다.

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] \quad (184)$$

이는 보다시피 데이터에 대하여 비선형이다. 여기에 강제로 (181)을 적용해보면 다음과 같다.

$$\hat{\sigma}^2 = \sum_{n=0}^{N-1} a_n x[n] \quad (185)$$

위 식의 기대값은 다음과 같다.

$$\mathbb{E}(\hat{\sigma}^2) = \sum_{n=0}^{N-1} a_n \mathbb{E}(x[n]) = 0 \quad (186)$$

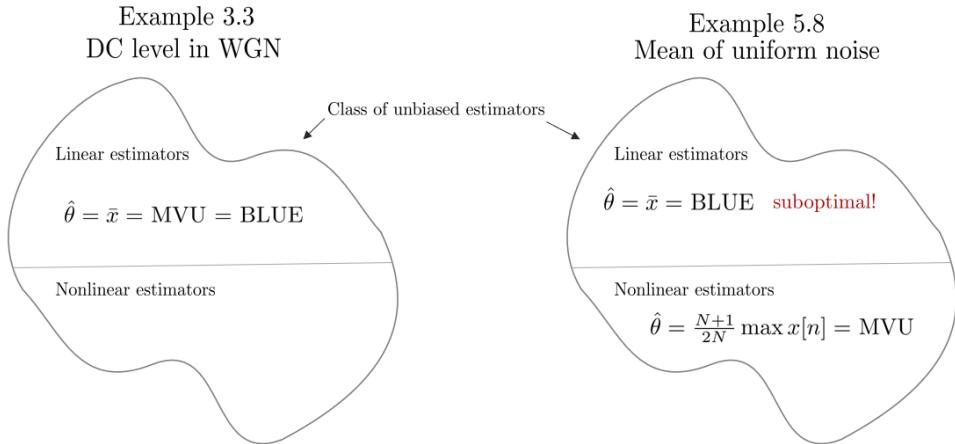
모든 n 에 대하여 $\mathbb{E}(x[n]) = 0$ 의 값을 가진다. 여기에서 우리는 단 하나의 선형 추정값도 찾을 수 없다. 비록 BLUE를 이러한 문제에 바로 적용하는 것은 어렵지만 데이터를 $y[n] = x^2[n]$ 과 같이 변형하게 되면 BLUE를 활용할 수 있는 여지가 생긴다.

$$\hat{\sigma}^2 = \sum_{n=0}^{N-1} a_n y[n] = \sum_{n=0}^{N-1} a_n x^2[n] \quad (187)$$

위 식의 기대값은 다음과 같다.

$$\mathbb{E}(\hat{\sigma}^2) = \sum_{n=0}^{N-1} a_n \sigma^2 = \sigma^2 \quad (188)$$

위 제약조건을 만족하는 다양한 a_n 의 값을 찾을 수 있다. **따라서 BLUE를 잘 사용하려면 데이터를 적절하게 잘 변환해야 한다.**



6.2 Finding the BLUE

BLUE를 찾기 위해서는 찾고자 하는 추정값 $\hat{\theta}$ 이 데이터에 선형 결합이면서 동시에 불편추정값(unbiased estimator)으로 제한한다. 다음으로 최소의 분산을 가지는 a_n 계수 값을 결정해야 한다. 불편추정값을 만족하기 위한 제약조건은 다음과 같다.

$$\mathbb{E}(\hat{\theta}) = \sum_{n=0}^{N-1} a_n \mathbb{E}(x[n]) = \theta \quad (189)$$

$\hat{\theta}$ 의 분산은 다음과 같다.

$$\text{var}(\hat{\theta}) = \mathbb{E} \left[\left(\sum_{n=0}^{N-1} a_n x[n] - \mathbb{E} \left(\sum_{n=0}^{N-1} a_n x[n] \right) \right)^2 \right] \quad (190)$$

$\mathbf{a} = [a_0, a_1, \dots, a_{N-1}]^\top$ 이라고 하면 위 식은 다음과 같은 벡터 형태로 전개된다.

$$\begin{aligned}
\text{var}(\hat{\theta}) &= \mathbb{E}[(\mathbf{a}^\top \mathbf{x} - \mathbf{a}^\top \mathbb{E}(\mathbf{x}))^2] \\
&= \mathbb{E}[(\mathbf{a}^\top (\mathbf{x} - \mathbb{E}(\mathbf{x})))^2] \\
&= \mathbb{E}[\mathbf{a}^\top (\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^\top \mathbf{a}] \\
&= \mathbf{a}^\top \mathbf{C} \mathbf{a}
\end{aligned} \tag{191}$$

최적의 벡터 \mathbf{a} 값은 (189)의 제약조건을 만족하면서 (191)를 최소화함으로써 찾을 수 있다.

Assumption:

이를 진행하기 전에 $\mathbb{E}(x[n])$ 의 형태에 대해 다시 정의할 필요가 있다. (189) 제약조건을 만족시키기 위해 $\mathbb{E}(x[n])$ 는 θ 에 대한 선형결합이라고 가정할 수 있다.

$$\mathbb{E}(x[n]) = s[n]\theta \tag{192}$$

여기서 $s[n]$ 은 우리가 이미 알고 있는 값이다. 만약 $\mathbb{E}(x[n]) = \cos \theta$ 와 같이 주어지면 불편추정값 제약조건을 만족하지 못하게 된다.

$$\begin{aligned}
\sum_{n=0}^{N-1} a_n \cos \theta &= \theta \\
\text{and } a_n?
\end{aligned} \tag{193}$$

따라서 $\mathbb{E}(x[n])$ 는 반드시 θ 에 대한 선형결합으로 표현되어야 한다. $\mathbb{E}(x[n])$ 를 사용하여 $x[n]$ 을 다시 쓰면 다음과 같다.

$$x[n] = \mathbb{E}(x[n]) + [x[n] - \mathbb{E}(x[n])] \tag{194}$$

여기서 $[x[n] - \mathbb{E}(x[n])]$ 는 노이즈 $w[n]$ 을 의미하므로 이는 다음과 같이 쓸 수 있다.

$$x[n] = \theta s[n] + w[n] \tag{195}$$

(192)와 같은 가정은 노이즈가 포함된 신호 크기 추정 문제에 BLUE를 적용할 수 있도록 해준다. 지금까지 내용을 정리해보자. BLUE를 찾기 위해서는 다음과 같은 분산을 최소화시키는 \mathbf{a} 를 찾아야 한다.

$$\text{var}(\hat{\theta}) = \mathbf{a}^\top \mathbf{C} \mathbf{a} \tag{196}$$

위 식은 불편추정값 제약 조건 (189)를 만족하면서 최소화되어야 한다.

$$\begin{aligned}
\sum_{n=0}^{N-1} a_n \mathbb{E}(x[n]) &= \theta \\
\sum_{n=0}^{N-1} a_n s[n] \theta &= \theta \\
\sum_{n=0}^{N-1} a_n s[n] &= 1 \\
\mathbf{a}^\top \mathbf{s} &= 1
\end{aligned} \tag{197}$$

$$- \mathbf{s} = [s_0, s_1, \dots, s[N-1]]^\top$$

최적화 문제로 표현하면 다음과 같다.

$$\text{BLUE : } \arg \min (\text{var}(\hat{\theta}) = \mathbf{a}^\top \mathbf{C} \mathbf{a}) \quad \text{subject to } \mathbf{a}^\top \mathbf{s} = 1 \tag{198}$$

위 식의 최적해는 다음과 같이 유도된다. 자세한 내용은 Appendix 6A를 참조하면 된다.

$$\mathbf{a}_{\text{opt}} = \frac{\mathbf{C}^{-1} \mathbf{s}}{\mathbf{s}^\top \mathbf{C}^{-1} \mathbf{s}} \tag{199}$$

따라서 BLUE와 최소 분산은 다음과 같다.

$$\boxed{\begin{aligned}\hat{\theta} &= \frac{\mathbf{s}^\top \mathbf{C}^{-1} \mathbf{x}}{\mathbf{s}^\top \mathbf{C}^{-1} \mathbf{s}} \\ \text{var}(\hat{\theta}) &= \frac{1}{\mathbf{s}^\top \mathbf{C}^{-1} \mathbf{s}}\end{aligned}} \quad (200)$$

(192)로부터 $\mathbb{E}(\mathbf{x}) = \theta \mathbf{s}$ 임을 알 수 있기 때문에 BLUE는 아래와 같이 편향되지 않음(unbiased)이 증명된다.

$$\begin{aligned}\mathbb{E}(\hat{\theta}) &= \frac{\mathbf{s}^\top \mathbf{C}^{-1} \mathbb{E}(\mathbf{x})}{\mathbf{s}^\top \mathbf{C}^{-1} \mathbf{s}} \\ &= \frac{\mathbf{s}^\top \mathbf{C}^{-1} \theta \mathbf{s}}{\mathbf{s}^\top \mathbf{C}^{-1} \mathbf{s}} \\ &= \theta\end{aligned} \quad (201)$$

앞서 서문에서 언급하였듯이 BLUE는 pdf에 대해서 자세히 모르는 상황에서도 처음 두 개의 모멘트 값

- \mathbf{s} : scaled 평균
- \mathbf{C} : 공분산

만 알아도 결정할 수 있다.

6.2.1 Example 6.1 - DC Level in White Noise

다음과 같은 관측 데이터가 주어졌다고 하자.

$$x[n] = A + w[n] \quad (202)$$

- $w[n]$: white noise σ^2 를 가지는 노이즈 (가우시안이 아닐 수 있음)

위 문제에서 파라미터 A 를 추정하고자 한다. $w[n]$ 은 가우시안이 아닐 수 있기 때문에 white noise(=서로 독립적인 노이즈)라고 하더라도 통계적으로는 서로 종속적일 수 있다. 위 식에서 $\mathbb{E}(x[n]) = A$ 므로 $s[n] = 1$ 이 되어 $\mathbf{s} = 1$ 이 된다. 따라서 BLUE는 다음과 같이 구할 수 있다.

$$\begin{aligned}\hat{A} &= \frac{\mathbf{1}^\top \frac{1}{\sigma^2} \mathbf{I} \mathbf{x}}{\mathbf{1}^\top \frac{1}{\sigma^2} \mathbf{I} \mathbf{1}} \\ &= \frac{1}{N} \sum_{n=0}^{N-1} x[n] \\ &= \bar{x}\end{aligned} \quad (203)$$

분산은 다음과 같이 구할 수 있다.

$$\begin{aligned}\text{var}(\hat{A}) &= \frac{1}{\mathbf{1}^\top \frac{1}{\sigma^2} \mathbf{1}} \\ &= \frac{\sigma^2}{N}\end{aligned} \quad (204)$$

따라서 pdf의 특성과 관계없이 BLUE는 데이터의 평균 \bar{x} 로 결정됨을 알 수 있다. 그리고 pdf가 가우시안 분포인 경우 BLUE는 MVUE가 된다.

6.3 Extension to a Vector Parameter

이번 섹션에서는 추정하고자 하는 파라미터가 $p \times 1$ 크기의 벡터 파라미터인 경우에 대해 알아본다.

$$\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n] \quad i = 1, 2, \dots, p \quad (205)$$

이를 행렬 형태로 나타내면 다음과 같다.

$$\hat{\theta} = \mathbf{Ax} \quad (206)$$

- $\mathbf{A} \in \mathbb{R}^{p \times N}$

$\hat{\theta}$ 가 불편추정값이기 위한 제약조건은 다음과 같다.

$$\mathbb{E}(\hat{\theta}_i) = \sum_{n=0}^{N-1} a_{in} \mathbb{E}(x[n]) = \theta_i \quad i = 1, 2, \dots, p \quad (207)$$

행렬 형태로 표현하면 다음과 같다.

$$\mathbb{E}(\hat{\theta}) = \mathbf{A}\mathbb{E}(\mathbf{x}) = \boldsymbol{\theta} \quad (208)$$

위의 불편추정값 제약조건이 충족되어야만 선형 추정값을 구할 수 있다는 것을 기억하자. $\mathbb{E}(\mathbf{x})$ 는 다음과 같은 형태여야 한다.

$$\mathbb{E}(\mathbf{x}) = \mathbf{H}\boldsymbol{\theta} \quad (209)$$

$$\mathbf{H} = \underbrace{\begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix}}_{\mathbf{H}} \boldsymbol{\theta} \quad (210)$$

(209)를 (208)에 대입하면 다음과 같은 식을 얻는다.

$$\mathbf{AH} = \mathbf{I} \quad (211)$$

i 번째 열벡터 $\mathbf{a}_i = [a_{i0}, a_{i1}, \dots, a_{i(N-1)}]^T$ 라고 하면 (206)는 $\hat{\theta}_i = \mathbf{a}_i^T \mathbf{x}_i$ 와 같이 쓸 수 있다. 행렬 \mathbf{A} 내부를 벡터 형태로 다시 쓰면 다음과 같다.

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_p^T \end{bmatrix} \quad (212)$$

\mathbf{H} 행렬의 i 번째 열벡터를 \mathbf{h}_i 라고 하면 이는 다음과 같이 쓸 수 있다.

$$\mathbf{H} = [\mathbf{h}_1 \quad \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_p] \quad (213)$$

따라서 벡터 곱으로 (211)를 다시 쓰면 아래와 같은 형태가 된다.

$$\mathbf{AH} = \mathbf{a}_i^T \mathbf{h}_j = \delta_{ij} \quad i = 1, 2, \dots, p; j = 1, 2, \dots, p \quad (214)$$

분산은 다음과 같이 쓸 수 있다.

$$\text{var}(\hat{\theta}_i) = \mathbf{a}_i^T \mathbf{C} \mathbf{a}_i \quad (215)$$

이전 섹션에서 스칼라 파라미터 케이스의 BLUE (198)와 동일하게 벡터 파라미터의 경우도 최적화 수식으로 나타낼 수 있다.

$$\boxed{\text{BLUE} : \arg \min (\text{var}(\hat{\theta}_i) = \mathbf{a}_i^T \mathbf{C} \mathbf{a}_i) \quad \text{subject to } \mathbf{a}_i^T \mathbf{h}_j = \delta_{ij}} \quad (216)$$

위 식을 만족하는 BLUE와 최소 분산은 다음과 같다. 자세한 유도 과정은 Appendix 6B를 참조하면 된다.

$$\boxed{\hat{\theta} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} \quad \mathbf{C}_{\hat{\theta}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}} \quad (217)$$

6.3.1 Theorem 6.1 (Gauss-Markov Theorem)

데이터가 아래와 같이 파라미터 $\boldsymbol{\theta}$ 에 대한 선형 모델 형태로 주어졌다고 가정하자.

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \quad (218)$$

- $\mathbf{H} \in \mathbb{R}^{N \times p}$
- $\boldsymbol{\theta} \in \mathbb{R}^{p \times 1}$
- $\mathbf{w} \in \mathbb{R}^{N \times 1}$: 평균이 0이고 공분산 \mathbf{C} 를 가지는 노이즈. (가우시안이 아닐 수 있음)

위 선형 모델에서 θ 에 대한 BLUE는 다음과 같다.

$$\hat{\theta} = (\mathbf{H}^\top \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{C}^{-1} \mathbf{x} \quad (219)$$

그리고 최소 분산은 다음과 같다.

$$\text{var}(\hat{\theta}_i) = [(\mathbf{H}^\top \mathbf{C}^{-1} \mathbf{H})^{-1}]_{ii} \quad (220)$$

마지막으로 공분산 행렬은 다음과 같다.

$$\mathbf{C}_{\hat{\theta}} = (\mathbf{H}^\top \mathbf{C}^{-1} \mathbf{H})^{-1} \quad (221)$$

7 Maximum Likelihood Estimation

이번 챕터에서는 MVUE가 존재하지 않거나 존재하더라도 찾을 수 없는 경우 이에 대한 대안을 찾는 방법에 대해 배운다. 대안으로 사용할 수 있는 추정값은 실제 추정 문제에서 매우 광범위하게 사용되는 maximum likelihood(ML) 원리를 기반으로 한다. ML 원리를 사용하면 아무리 복잡한 문제에서도 (데이터가 충분히 주어졌다면) 효율적으로 추정값을 구할 수 있다. 또한 ML을 사용한 추정값은 근사적으로 efficiency한 특성이 있기 때문에 근사적으로 MVUE를 만족한다. 이러한 이유 때문에 실제 추정 문제에서는 광범위하게 ML 원리를 사용하여 추정값을 구하고 있다. 이렇게 ML 원리를 기반으로 하는 추정 방법을 maximum likelihood estimation(MLE)라고 한다.

7.1 An Example

이번 섹션에서는 MVUE를 찾기 어려운 예제를 살펴보자. MVUE를 구할 수 없는 경우에도 MLE를 사용하면 점근적으로(asymptotically) efficient한 특성을 지니게 되고 따라서 근사적인 MVUE를 찾을 수 있다는 것을 보일 것이다.

7.1.1 Example 7.1 - DC Level in White Gaussian Noise - Modified

다음과 같은 관측 데이터가 주어졌다고 하자.

$$x[n] = A + w[n] \quad (222)$$

- $w[n] \sim \mathcal{N}(0, A)$: WGN

찾고자 하는 파라미터 A 는 미지의 값이며 DC Level이기 때문에 양수 $A > 0$ 라고 가정할 수 있다. 이번 예제에서 $w[n]$ 은 WGN이며 분산 A 를 가진다고 하자. 이는 지금까지 다른 Example 3.3과 같은 일반적인 예제의 경우와는 다르게 A 값이 평균과 분산에 모두 영향을 미친다.

Trial 1 (CRLB):

MVUE를 찾기 위해 우선 CRLB를 만족하는지 알아봐야 한다. CRLB의 정규 조건(regularity condition)을 알아보기 위해 pdf를 전개하면 다음과 같다.

$$p(\mathbf{x}; A) = \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp \left[-\frac{1}{2A} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \quad (223)$$

로그 가능도함수로 변경하고 미분 후 다음과 같다.

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x}; A)}{\partial A} &= -\frac{N}{2A} + \frac{1}{A} \sum_{n=0}^{N-1} (x[n] - A) + \frac{1}{2A^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \\ &\stackrel{?}{=} I(A)(\hat{A} - A) \end{aligned} \quad (224)$$

위 식을 통해 CRLB 정규 조건을 만족하지 않는다는 것을 알 수 있고 따라서 efficient한 추정값이 존재하지 않는다는 것을 알 수 있다. CRLB를 사용하여 MVUE를 구할 수 없다는 의미이다. 따라서 다음을 만족하는 \hat{A} 를 찾아야 한다.

$$\text{var}(\hat{A}) \geq \frac{A^2}{N(A + \frac{1}{2})} \quad (225)$$

Trial 2 (RBLS - Simple Ver):

다음으로 충분통계량(sufficient statistic)을 사용하여 MVUE를 찾아보자. (223)의 exponential 내부 항을 전개하면 다음과 같다.

$$\frac{1}{A} \sum_{n=0}^{N-1} (x[n] - A)^2 = \frac{1}{A} \sum_{n=0}^{N-1} x[n] - 2N\bar{x} + NA \quad (226)$$

Neymann-Fisher factorization에 의해 pdf는 다음과 같이 분해할 수 있다.

$$p(\mathbf{x}; A) = \underbrace{\frac{1}{(2\pi A)^{\frac{N}{2}}} \exp \left[-\frac{1}{2} \left(\frac{1}{A} \sum_{n=0}^{N-1} x^2[n] + NA \right) \right]}_{g\left(\sum_{n=0}^{N-1} x^2[n]; A\right)} \underbrace{\exp(N\bar{x})}_{h(\mathbf{x})} \quad (227)$$

Neymann-Fisher factorization에 따라 파라미터 A 에 대한 충분통계량은 $T(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n]$ 임을 알 수 있다. 다음 스텝은 충분통계량에 대한 유일한 함수 $g(T(\mathbf{x}))$ 를 찾음으로써 충분통계량이 complete한지 여부를 판단하는 것이다. $g(\cdot)$ 은 다음 수식을 만족해야 한다.

$$\mathbb{E} \left[g \left(\sum_{n=0}^{N-1} x^2[n] \right) \right] = A \quad \text{for all } A > 0 \quad (228)$$

위 식의 $\mathbb{E}[g(\cdot)]$ 를 찾기 위해 앞서 $\mathbb{E} \left[\sum_{n=0}^{N-1} x^2[n] \right]$ 를 전개해보면 다음과 같다.

$$\begin{aligned} \mathbb{E} \left[\sum_{n=0}^{N-1} x^2[n] \right] &= N \mathbb{E}[x^2[n]] \\ &= N[\text{var}(x[n]) + \mathbb{E}^2(x[n])] \\ &= N(A + A^2) \end{aligned} \quad (229)$$

위 식 (229)에서 오직 A 만 남기고 나머지 bias를 제거할 수 있는 형태의 $g(\cdot)$ 를 찾을 수 없다. 이는 Example 5.8 처럼 단순하게 스케일 값 $1/N$ 만 곱해줘서 해결되지 않는다. 따라서 충분통계량은 complete하지 않고 MVUE를 구할 수 없다.

Trial 3 (RBLS - Complicated Ver):

RBLS의 조금 더 복잡한 방법으로는 임의의 불편추정값 \hat{A} 에 대하여 $\mathbb{E}(\hat{A} | \sum_{n=0}^{N-1} x^2[n])$ 와 같은 조건부 pdf를 구함으로써 MVUE 추정값을 얻을 수 있었다. $\hat{A} = x[0]$ 로 설정하면 기대값은 다음과 같다.

$$\mathbb{E}(x[0] | \sum_{n=0}^{N-1} x^2[n]) \quad (230)$$

하지만 이를 만족하는 조건부 pdf를 구하는 일은 만만치 않고 앞서 보았듯이 충분통계량이 complete하지 않기 때문에 적절한 추정값을 얻을 수 없다.

Trial 4 :

지금까지 최적의 추정값을 얻기 위한 여러 방법을 시도해보았으나 적절한 추정값을 구할 수 없었다. 차선의 방법은 \hat{A} 를 일단 평균으로 설정해보는 것이다.

$$\hat{A}_1 = \begin{cases} \bar{x} & \text{if } \bar{x} > 0 \\ 0 & \text{if } \bar{x} \leq 0 \end{cases} \quad (231)$$

우리는 DC Level이 $A > 0$ 인 사실을 알기 때문에 위와 같이 추정한다. 다음 방법으로는 \hat{A} 를 분산으로 추정해보는 것이다.

$$\hat{A}_2 = \frac{1}{N-1} \sum_{n=0}^{N-1} (x[n] - \hat{A}_1)^2 \quad (232)$$

하지만 위 두 추정값은 최적의 추정값이라는 사실을 어디에서도 보장받을 수 없다.

위 예제와 같이 MVUE를 정확하게 구할 수 없는 경우에는 근사적으로라도 최적의 추정값을 찾아야 한다. 만약 데이터가 $N \rightarrow \infty$ 와 같이 무한히 큰 경우 추정값은 점근적으로 efficient해진다.

$$\begin{aligned}\mathbb{E}(\hat{A}) &\rightarrow A \\ \text{var}(\hat{A}) &\rightarrow \text{CRLB}\end{aligned}\tag{233}$$

CRLB는 (225)의 분산을 의미한다. 위 식의 첫번째 조건을 만족하는 추정값은 점근적으로 편향되지 않았다 (asymptotically unbiased)라고 할 수 있으며 두 번째 조건을 만족하는 추정값은 점근적으로 efficient하다 (asymptotically efficient)라고 할 수 있다. 하지만 현실 추정 문제에서는 데이터가 유한한 경우가 대부분이므로 이러한 조건들을 적용하는 것이 쉽지 않다.

7.2 Finding the MLE

7.2.1 Example 7.2 and 7.3 - DC Level in White Gaussian Noise - Modified (continued)

Maximum likelihood estimation(MLE)는 로그 가능도함수가 파라미터 A 에 대한 함수이고 동시에 exponential 항이 2차식(quadratic form)인 특성을 활용하여 미분 후 0이 되는 값을 찾는 방법이다. 미분 후 0인 값은 로그 가능도함수의 극대값이 되며 이 값이 곧 maximum likelihood가 된다. 로그 가능도 함수의 미분은 (224)와 같다.

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = -\frac{N}{2A} + \frac{1}{A} \sum_{n=0}^{N-1} (x[n] - A) + \frac{1}{2A^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \tag{234}$$

위 식을 0으로 설정하고 정리하면 다음과 같다.

$$\hat{A}^2 + \hat{A} - \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] = 0 \tag{235}$$

\hat{A} 에 대하여 위 식을 정리하면 다음과 같다.

$$\hat{A} = -\frac{1}{2} \pm \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}} \tag{236}$$

DC Level $A > 0$ 인 경우에 이 중 +인 항이 곧 솔루션이 된다.

$$\boxed{\hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}} \tag{237}$$

MLE 추정값의 기대값을 보면 다음과 같이 편향되어(biased) 있음을 알 수 있다.

$$\begin{aligned}\mathbb{E}(\hat{A}) &= \mathbb{E}\left(-\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}\right) \\ &\neq -\frac{1}{2} + \sqrt{\mathbb{E}\left(\frac{1}{N} \sum_{n=0}^{N-1} x^2[n]\right) + \frac{1}{4}} \quad \text{for all } A \\ &= -\frac{1}{2} + \sqrt{A + A^2 + \frac{1}{4}} \\ &= A\end{aligned}\tag{238}$$

$$\therefore \mathbb{E}(\hat{A}) \neq A \tag{239}$$

만약 데이터가 $N \rightarrow \infty$ 와 같이 무한히 크다면 다음과 같이 MLE는 점근적으로 편향성이 없어진다.

$$\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] \rightarrow \mathbb{E}(x^2[n]) = A + A^2 \tag{240}$$

(237)는 다음과 같아진다.

$$\therefore \hat{A} \rightarrow A \quad \text{for } N \rightarrow \infty \tag{241}$$

7.3 Properties of the MLE

7.3.1 Theorem 7.1 (Asymptotic Properties of the MLE)

만약 pdf $p(\mathbf{x}; \theta)$ 가 정규 조건(regular condition)을 만족한다면 미지의 파라미터 θ 에 대한 MLE 추정값은 점근적인(asymptotically) 가우시안 분포를 가진다.

$$\hat{\theta} \xrightarrow{a} \mathcal{N}(\theta, I^{-1}(\theta)) \quad (242)$$

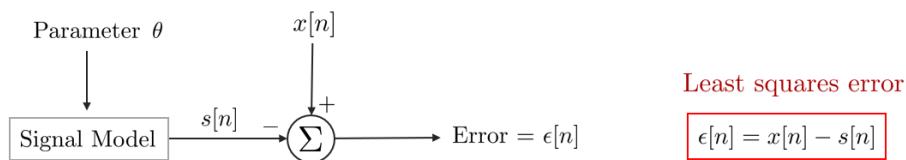
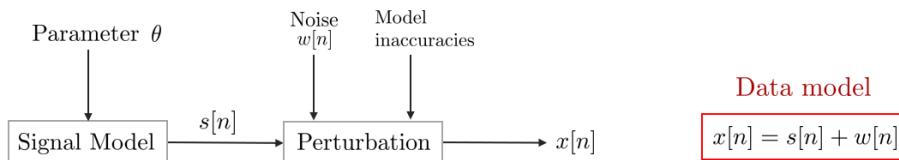
$I(\theta)$ 는 θ 에 대한 Fisher information 값을 의미한다. 정규 조건은 다음 조건을 만족해야 한다.

로그 가능도함수가 미분 가능해야 한다. Fisher information 값이 0이 아닌 값을 가져야 한다. 점근적인 가우시안 분포 성질에 따라 MLE는 점근적으로 efficient한 특성을 지니며 또한 점근적으로 MVUE를 만족한다. 실제 추정 문제에서 더 정확한 MLE 추정값을 얻기 위해서는 더 많은 데이터가 요구된다.

8 Least Squares

이전 챕터에서 우리는 최적 또는 최적에 근사한(데이터가 많은 경우) 추정값을 찾는 방법에 대해 학습하였다. 이러한 추정값들은 전부 편향되지 않은(unbiased) 추정값들의 집합에서 가장 분산이 작은 추정값을 찾음으로써 최종적으로 MVUE를 찾는 과정이었다. 이번 챕터에서는 이러한 철학과는 다소 다른 최소제곱법에 대해 학습한다. 최소제곱법은 1795년에 가우스가 행성의 운동을 연구하기 위해 사용한 방법이다. 이 방법의 중요한 특징은 데이터에 대한 확률적 가정이 이루어지지 않는다는 것이며 오직 신호 모델로만 가정된다. 최소제곱법의 추정값은 비록 통계적인 성능을 평가할 수 없지만 수 많은 실제 추정 문제에서 최소제곱법 추정값은 널리 사용되기 때문에 반드시 알아야 할 개념 중 하나이다.

8.1 The Least Squares Approach



최소제곱법(least squares, LS) 방법을 수학적으로 설명하면 주어진 데이터 $x[n]$ 과 노이즈가 없다고 가정하는 신호 $s[n]$ 가 있을 때 둘의 차이의 제곱을 최소화하는 문제라고 볼 수 있다. 여기서 신호 $s[n]$ 은 파라미터 θ 에 종속적인 모델로부터 생성된다. $s[n]$ 는 확률변수가 아닌 순수하게 주어진 데이터이다(deterministic). 즉, 확률로 모델링할 수 없다. 관측 노이즈와 모델의 부정확성 등으로 신호는 변질되어 우리는 변질된 신호 데이터 $x[n]$ 를 얻게 된다. 파라미터 θ 에 대한 least squares estimator(LSE)는 $s[n]$ 을 최대한 $x[n]$ 에 가까워지도록 θ 를 조정하는 역할을 한다. 가까운 정도는 다음과 같은 LS error criterion으로 측정된다.

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n])^2 \quad (243)$$

$$\hat{\theta}_{\text{LSE}} = \arg \min J(\theta) \quad (244)$$

위 식에서 $x[n]$ 은 기준까지 다른 확률변수가 아님에 유의한다. $s[n]$ 이 확률변수가 아니므로 따라서 $x[n]$ 또한 확률변수가 아니다. LSE는 가우시안 노이즈 뿐만 아니라 기타 다른 분포의 노이즈에 대해서도 잘 동작하지만 당연히 LSE의 성능은 노이즈의 종류 또는 크기에 종속적이다. LSE는 일반적으로 정확한 데이터의 확률적 특성을 알지 못하거나 최적의 추정값을 찾는 것이 매우 복잡하거나 불가능할 때 주로 사용된다.

8.1.1 Example 8.1 - DC Level Signal

(243)에서 $s[n] = A$ 인 경우를 생각해보자. LSE criterion은 다음과 같이 쓸 수 있다.

$$J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2 \quad (245)$$

LSE는 항상 2차식 형태(quadratic form)이므로 A 에 대하여 미분 후 0이 되는 값을 찾으면 그 값이 극소값이 된다. 위 식을 미분하면 다음과 같다.

$$\begin{aligned} \frac{\partial J(A)}{\partial A} &= -2 \sum_{n=0}^{N-1} (x[n] - A) = 0 \\ &= \sum_{n=0}^{N-1} x[n] - NA = 0 \end{aligned} \quad (246)$$

$$\begin{aligned} \hat{A} &= \frac{1}{N} \sum_{n=0}^{N-1} x[n] \\ &= \bar{x} \end{aligned}$$

(247)

이는 데이터의 평균이므로 만약 노이즈가 가우시안 분포를 따르는 경우 \hat{A} 는 MVUE가 된다.

8.2 Linear Least Squares

8.2.1 Scalar Case

선형 LS 방법을 사용하기 위해 신호 $s[n]$ 은 데이터 시퀀스 $h[n]$ 에 찾고자 하는 미지의 파라미터 θ 가 곱해진 값이라고 가정한다.

$$s[n] = \theta h[n] \quad (248)$$

위 식에서 $h[n]$ 은 이미 알고 있는 값이다. 다음으로 LSE criterion은 다음과 같다.

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - \theta h[n])^2 \quad (249)$$

앞선 예제와 동일하게 LSE는 항상 2차식 형태(quadratic form)이므로 이를 미분하여 0이 되는 값이 극소값이 된다. 미분을 수행하면 다음과 같다.

$$\frac{\partial J(\theta)}{\partial \theta} = -2 \sum_{n=0}^{N-1} h[n](x[n] - \theta h[n]) = 0 \quad (250)$$

따라서 아래와 같은 LSE $\hat{\theta}$ 를 얻는다.

$$\hat{\theta} = \frac{\sum_{n=0}^{N-1} x[n]h[n]}{\sum_{n=0}^{N-1} h^2[n]}$$

(251)

(251)를 (249)에 대입하면 다음과 같은 최소 LS 에러값 J_{\min} 을 얻는다.

$$\begin{aligned} J_{\min} = J(\hat{\theta}) &= \sum_{n=0}^{N-1} (x[n] - \hat{\theta}h[n])(x[n] - \hat{\theta}h[n]) \\ &= \sum_{n=0}^{N-1} x[n](x[n] - \hat{\theta}h[n]) - \hat{\theta} \underbrace{\sum_{n=0}^{N-1} h[n](x[n] - \hat{\theta}h[n])}_{S \rightarrow 0} \\ &= \sum_{n=0}^{N-1} x^2[n] - \hat{\theta} \sum_{n=0}^{N-1} x[n]h[n] \end{aligned} \quad (252)$$

위 식에 $\hat{\theta}$ 를 대입하면 S 부분은 0가 되어 사라진다. 따라서 J_{\min} 은 다음과 같다.

$$J_{\min} = \sum_{n=0}^{N-1} x^2[n] - \frac{\left(\sum_{n=0}^{N-1} x[n]h[n] \right)^2}{\sum_{n=0}^{N-1} h^2[n]} \quad (253)$$

최소 LS 에러값은 관측 데이터의 제곱값 $\sum_{n=0}^{N-1} x^2[n]$ 보다 항상 작거나 같은 값을 가진다.

$$0 \leq J_{\min} \leq \sum_{n=0}^{N-1} x^2[n] \quad (254)$$

8.2.2 Vector Case

다음으로 추정하고자 하는 파라미터가 벡터 파라미터 $\theta \in \mathbb{R}^{p \times 1}$ 인 경우에 대해 알아보자. 신호는 $s = [s[0], s[1], \dots, s[N-1]]^\top$ 과 같이 나타낼 수 있고 (248)와 같이 신호는 데이터 시퀀스와 파라미터 θ 의 곱으로 표현한다.

$$\mathbf{s} = \mathbf{H}\theta \quad (255)$$

\mathbf{H} 는 $N \times p$ 크기의 행렬($N > p$)이며 rank가 p 인 full rank 행렬이다. 일반적으로 \mathbf{H} 는 관측 행렬(observation matrix)라고 부른다. LSE criterion은 다음과 같이 쓸 수 있다.

$$\begin{aligned} J(\theta) &= \sum_{n=0}^{N-1} (x[n] - s[n])^2 \\ &= (\mathbf{x} - \mathbf{H}\theta)^\top (\mathbf{x} - \mathbf{H}\theta) \end{aligned} \quad (256)$$

스칼라 파라미터 케이스와 동일하게 위 식을 미분 후 0이 되는 값이 극소값이다. 우선 $J(\theta)$ 를 전개하면 다음과 같다.

$$\begin{aligned} J(\theta) &= \mathbf{x}^\top \mathbf{x} - \mathbf{x}^\top \mathbf{H}\theta - \theta^\top \mathbf{H}^\top \mathbf{x} + \theta^\top \mathbf{H}^\top \mathbf{H}\theta \\ &= \mathbf{x}^\top \mathbf{x} - 2\mathbf{x}^\top \mathbf{H}\theta + \theta^\top \mathbf{H}^\top \mathbf{H}\theta \end{aligned} \quad (257)$$

위 식을 θ 에 대하여 미분한다.

$$\frac{\partial J(\theta)}{\partial \theta} = -2\mathbf{H}^\top \mathbf{x} + 2\mathbf{H}^\top \mathbf{H}\theta = 0 \quad (258)$$

LSE $\hat{\theta}$ 는 다음과 같다.

$$\boxed{\hat{\theta} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x}} \quad (259)$$

위 식을 정규 방정식(normal equation)의 해라고 한다. 행렬 \mathbf{H} 를 full rank로 가정함으로써 $\mathbf{H}^\top \mathbf{H}$ 가 역행렬이 존재함을 보장하였다. 놀랍게도 LSE는 CRLB를 통해 구한 efficient 추정값과 BLUE를 통해 구한 추정값과 동일한 형태를 가진다. 최소 LS 에러값은 다음과 같이 $\hat{\theta}$ 을 대입함으로써 얻을 수 있다.

$$\begin{aligned} J_{\min} &= J(\hat{\theta}) \\ &= (\mathbf{x} - \mathbf{H}\hat{\theta})^\top (\mathbf{x} - \mathbf{H}\hat{\theta}) \\ &= (\mathbf{x} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x})^\top (\mathbf{x} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x}) \\ &= \mathbf{x}^\top (\mathbf{I} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top) (\mathbf{I} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top) \mathbf{x} \\ &= \mathbf{x}^\top (\mathbf{I} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top) \mathbf{x} \end{aligned} \quad (260)$$

위 식의 네 번째 줄에서 $(\mathbf{I} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top)$ 항은 멱동 행렬(idempotent matrix)이므로 $\mathbf{A}^2 = \mathbf{A}$ 의 특성을 지닌다. 따라서 둘 중 하나는 소거되어 다섯번 째 줄이 유도된다. 다른 형태로 유도한 최소 LS 에러값은 다음과 같다.

$$\begin{aligned} J_{\min} &= J(\hat{\theta}) \\ &= \mathbf{x}^\top \mathbf{x} - \mathbf{x}^\top \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x} \\ &= \mathbf{x}^\top (\mathbf{x} - \mathbf{H}\hat{\theta}) \end{aligned} \quad (261)$$

8.2.3 Vector Weighted Case

다음으로 LSE criterion에 가중치가 곱해져 있는 경우에 대해 학습한다. 이를 일반적으로 weighted least squares(WLS)라고 한다. LSE criterion 중간에 $N \times N$ 크기의 positive definite이면서 대칭인 \mathbf{W} 행렬이 곱해진다.

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top \mathbf{W} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \quad (262)$$

만약 \mathbf{W} 가 대각행렬이고 대각 성분이 $[\mathbf{W}]_{ii} = w_i > 0$ 인 경우 LS 에러값은 다음과 같이 쓸 수 있다.

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} w_n (x[n] - A)^2 \quad (263)$$

위 식은 $x[n] = A + w[n]$ 이고 $w[n] \sim \mathcal{N}(0, \sigma^2)$ 인 데이터가 주어졌을 때 가중치 행렬의 값들을 $w_n = 1/\sigma^2$ 와 같이 선택한 것과 동일하다. 미분 후 0인 값을 찾으면 다음과 같은 LSE \hat{A} 를 얻을 수 있다.

$$\begin{aligned} \hat{A} &= \frac{\sum_{n=0}^{N-1} w[n]x[n]}{\sum_{n=0}^{N-1} w[n]} \\ &= \frac{\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}} \end{aligned} \quad (264)$$

따라서 weighted LSE \hat{A} 는 노이즈 $w[n]$ 가 white noise 인 경우($\mathbf{W} = \mathbf{C}^{-1}$) BLUE (200)와 동일한 추정값을 가지는 것을 알 수 있다. 벡터 형태로 표현한 일반적인 LSE $\hat{\boldsymbol{\theta}}$ 와 최소 LS 에러값 J_{\min} 은 다음과 같다.

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{W} \mathbf{x} \quad (265)$$

$$J_{\min} = \mathbf{x}^\top (\mathbf{W} - \mathbf{W} \mathbf{H} (\mathbf{H}^\top \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{W}) \mathbf{x} \quad (266)$$

8.3 Geometrical Interpretations

이번 섹션에서는 LS를 기하학적 관점에서 해석하는 방법에 대해 설명한다. 기하학적 설명을 통해 수식의 유도 과정을 보다 직관적으로 이해할 수 있으며 추가적으로 유용한 성질들을 도출해낼 수 있다. 일반적인 신호 모델 $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$ 이 주어졌을 때 이를 열벡터(column vector) 형태로 표현하면 다음과 같다.

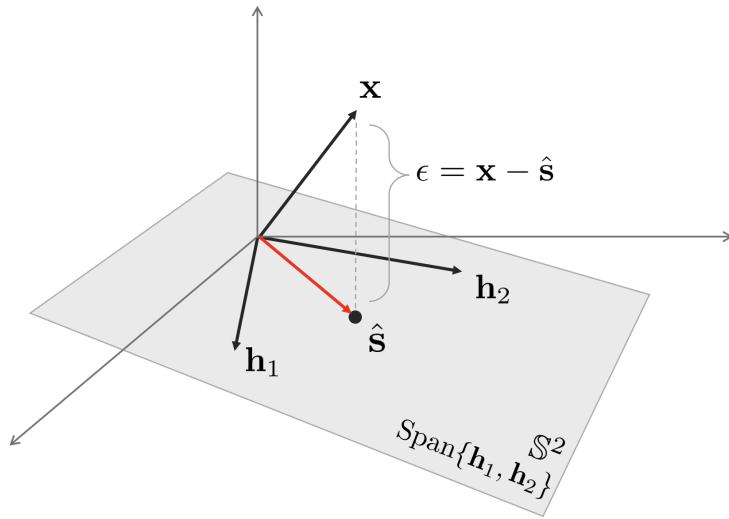
$$\begin{aligned} \mathbf{s} &= [\mathbf{h}_1 \quad \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_p] \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \\ &= \sum_{i=1}^p \theta_i \mathbf{h}_i \end{aligned} \quad (267)$$

위 식에서 보다시피 신호 모델은 각 신호 벡터 $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_p\}$ 들의 선형 결합(linear combination)으로 볼 수 있다. LSE criterion (256)은 다음과 같이 쓸 수 있다.

$$\begin{aligned} J(\boldsymbol{\theta}) &= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \\ &= \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 \\ &= \left\| \mathbf{x} - \sum_{i=1}^p \theta_i \mathbf{h}_i \right\|^2 \end{aligned} \quad (268)$$

위 식에서 선형 LS 방법은 데이터 \mathbf{x} 와 신호 벡터 $\sum_{i=1}^p \theta_i \mathbf{h}_i$ 의 차이의 제곱을 최소화하는 것으로 볼 수 있다. 기하학적으로 해석하자면 데이터 \mathbf{x} 는 N 차원의 벡터 공간 \mathbb{R}^N 에 존재하는 반면, 신호 벡터는 $p < N$ 인 subspace \mathbb{S}^p 에 대한 공간 \mathbb{S}^p 에 존재한다. (\mathbf{H} 는 full rank라는 가정으로 인해 모든 열벡터들은 독립이며 따라서 p 차원의 subspace 를 스팬한다고 볼 수 있다.)

$N = 3, p = 2$ 인 경우를 생각해보자. $p = 2$ 이므로 $\mathbf{H} = [\mathbf{h}_1^\top, \mathbf{h}_2^\top]^\top$ 과 이며 찾고자 하는 파라미터는 θ_1, θ_2 가 된다. 아래 그림과 같이 $\mathbf{h}_1, \mathbf{h}_2$ 벡터에 의해 \mathbb{S}^2 subspace가 스팬된다.



$N = 3$ 이기 때문에 주어진 데이터 \mathbf{x} 는 S^2 공간에 존재하지 않는다. 직관적으로 보면 알 수 있듯이 \mathbf{x} 와 S^2 공간 사이의 최소 거리는 \mathbf{x} 에서 수선의 발을 내린 $\hat{\mathbf{s}}$ 가 된다. 이를 S^2 공간에 대한 \mathbf{x} 의 직교 프로젝션(orthogonal projection)이라고 한다. 이는 여러 벡터 ϵ 이 S^2 에 존재하는 모든 벡터에 대하여 직교한다는 것을 의미한다. 두 벡터 \mathbf{a}, \mathbf{b} 가 직교한다는 의미는 $\mathbf{a}^\top \mathbf{b}$ 와 동치이므로 다음 공식이 성립한다.

$$(\mathbf{x} - \hat{\mathbf{s}}) \perp S^2 \quad (269)$$

이는 $\mathbf{h}_1, \mathbf{h}_2$ 에 대해서도 성립한다.

$$\begin{aligned} (\mathbf{x} - \hat{\mathbf{s}}) &\perp \mathbf{h}_1 \\ (\mathbf{x} - \hat{\mathbf{s}}) &\perp \mathbf{h}_2 \end{aligned} \quad (270)$$

내적의 성질에 의해 다음 공식이 성립한다.

$$\begin{aligned} (\mathbf{x} - \hat{\mathbf{s}})^\top \mathbf{h}_1 &= 0 \\ (\mathbf{x} - \hat{\mathbf{s}})^\top \mathbf{h}_2 &= 0 \end{aligned} \quad (271)$$

$\hat{\mathbf{s}} = \theta_1 \mathbf{h}_1 + \theta_2 \mathbf{h}_2$ 이므로 이를 대입하면 다음과 같다.

$$\begin{aligned} (\mathbf{x} - \theta_1 \mathbf{h}_1 + \theta_2 \mathbf{h}_2)^\top \mathbf{h}_1 &= 0 \\ (\mathbf{x} - \theta_1 \mathbf{h}_1 + \theta_2 \mathbf{h}_2)^\top \mathbf{h}_2 &= 0 \end{aligned} \quad (272)$$

행렬 형태로 취합하면 다음과 같다.

$$\begin{aligned} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top \mathbf{h}_1 &= 0 \\ (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top \mathbf{h}_2 &= 0 \end{aligned} \quad (273)$$

두 식을 하나의 식으로 통합하면 다음과 같다.

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top [\mathbf{h}_1, \mathbf{h}_2] = \mathbf{0}^\top \quad (274)$$

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top \mathbf{H} = \mathbf{0}^\top \quad (275)$$

따라서 (259)와 동일한 LSE $\hat{\boldsymbol{\theta}}$ 를 구할 수 있다.

$$\boxed{\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x}} \quad (276)$$

만약 여러 벡터를 $\epsilon = \mathbf{x} - \mathbf{H}\boldsymbol{\theta}$ 라고 하면 LSE는 다음 식을 통해 바로 구할 수 있다.

$$\epsilon^\top \mathbf{H} = \mathbf{0}^\top \quad (277)$$

위 식과 같이 여러 벡터 ϵ 은 반드시 \mathbf{H} 의 열벡터와 직교해야 한다. 이를 직교성 원리(orthogonal principle)이다. 앞서 구한 $\hat{\boldsymbol{\theta}}$ 를 신호 모델에 대입하면 다음과 같다.

$$\begin{aligned}
\hat{\mathbf{s}} &= \mathbf{H}\hat{\boldsymbol{\theta}} \\
&= \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x} \\
&= \mathbf{P}\mathbf{x}
\end{aligned} \tag{278}$$

위 식에서 $\mathbf{P} = \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \in \mathbb{R}^{N \times N}$ 행렬을 직교 프로젝션 행렬(orthogonal projection matrix)라고 한다. \mathbf{P} 는 다음과 같은 성질을 가지고 있다.

- $\mathbf{P}^\top = \mathbf{P}$: 대칭(symmetric) 행렬
- $\mathbf{P}^2 = \mathbf{P}$: 멱동(idempotent) 행렬

\mathbf{P} 행렬을 사용하여 에러 벡터 ϵ 을 표현하면 다음과 같다.

$$\begin{aligned}
\epsilon &= \mathbf{x} - \hat{\mathbf{s}} \\
&= \mathbf{x} - \mathbf{P}\mathbf{x} \\
&= (\mathbf{I} - \mathbf{P})\mathbf{x} \\
&= \mathbf{P}^\perp \mathbf{x}
\end{aligned} \tag{279}$$

$\mathbf{P}^\perp = \mathbf{I} - \mathbf{P}$ 또한 \mathbf{P} 와 동일한 성질을 지닌 프로젝션 행렬이다. 최종적으로 최소 LS 에러값 J_{\min} 은 다음과 같다.

$$\begin{aligned}
J_{\min} &= J(\hat{\boldsymbol{\theta}}) \\
&= (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^\top (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) \\
&= (\mathbf{x} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x})^\top (\mathbf{x} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x}) \\
&= \mathbf{x}^\top (\mathbf{I} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top)(\mathbf{I} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top)\mathbf{x} \\
&= \mathbf{x}^\top (\mathbf{I} - \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top)\mathbf{x} \\
&= \mathbf{x}^\top (\mathbf{I} - \mathbf{P})\mathbf{x} \\
&= \mathbf{x}^\top \mathbf{P}^\perp \mathbf{x} \\
&= \mathbf{x}^\top \mathbf{P}^\perp \mathbf{x} \\
&= \|\mathbf{P}^\perp \mathbf{x}\|^2 \\
&= \|\epsilon\|^2
\end{aligned} \tag{280}$$

8.4 Sequential Least Squares

많은 신호 처리 문제에서 데이터를 받을 때 연속 시간에 대한 신호를 샘플링하여 받고 있다. 지금까지 배운 LSE는 최적의 추정값을 제공하지만 새로운 신호가 올 때마다 모든 데이터에 대한 정규 방정식 (259)을 새로 계산해야 한다. 이번 섹션에서는 $\mathbf{x} = \{x[1], x[2], \dots, x[N-1]\}$ 이 주어진 상태에서 새로운 데이터 $x[n]$ 이 들어 왔을 때 정규 방정식을 푸는 해법이 아닌 순차적(sequential)으로 최적의 추정값 LSE $\hat{\boldsymbol{\theta}}$ 를 업데이트하는 방법에 대해 배운다.

8.4.1 Sequential LS (scalar parameter)

다음과 같은 DC Level A 파라미터 추정 문제가 주어졌다고 가정하자.

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1 \tag{281}$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

$\hat{A}[N-1]$ 은 다음과 같이 LSE를 사용하여 구할 수 있다.

$$\hat{A}[N-1] = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \tag{282}$$

이 상태에서 새로운 데이터 $x[N]$ 이 들어왔다고 하자. LSE는 다음과 같이 된다.

$$\begin{aligned}
\hat{A}[N] &= \frac{1}{N+1} \sum_{n=0}^N x[n] \\
&= \frac{1}{N+1} \left(\sum_{n=0}^{N-1} x[n] + x[N] \right) \\
&= \frac{N}{N+1} \hat{A}[N-1] + \frac{1}{N+1} x[N]
\end{aligned} \tag{283}$$

위 식에서 보다시피 새로운 데이터 $x[n]$ 에 대한 LSE $\hat{A}[N]$ 는 이전 LSE $\hat{A}[N - 1]$ 로부터 구할 수 있다. 위 식을 정리하여 다시쓰면 아래와 같다.

$$\hat{A}[N] = \hat{A}[N - 1] + \underbrace{\frac{1}{N+1}(x[N] - \hat{A}[N - 1])}_{\text{correction}} \quad (284)$$

새로운 LSE 항은 이전 LSE 항에 correction 항이 추가된 것으로 볼 수 있다. 최소 LS 에러값 J_{\min} 에 위 식을 넣고 전개하면 다음과 같이 재귀적인 식으로 변형된다.

$$J_{\min}[N - 1] = \sum_{n=0}^{N-1} (x[n] - \hat{A}[N - 1])^2 \quad (285)$$

$$J_{\min}[N] = \sum_{n=0}^N (x[n] - \hat{A}[N])^2 \quad (286)$$

위 식 $\hat{A}[N]$ 에 (284)를 넣고 전개하면 다음과 같다.

$$\begin{aligned} J_{\min}[N] &= \sum_{n=0}^{N-1} \left[x[n] - \hat{A}[N - 1] - \frac{1}{N+1}(x[N] - \hat{A}[N - 1]) \right]^2 + (x[N] - \hat{A}[N])^2 \\ &= J_{\min}[N - 1] - \frac{2}{N+1} \sum_{n=0}^{N-1} (x[n] - \hat{A}[N - 1])(x[N] - \hat{A}[N - 1]) \\ &\quad + \frac{N}{(N+1)^2}(x[N] - \hat{A}[N - 1])^2 + (x[N] - \hat{A}[N])^2 \end{aligned} \quad (287)$$

정리하면 최소 LS 에러값에 대한 재귀식을 얻을 수 있다.

$$J_{\min}[N] = J_{\min}[N - 1] + \frac{N}{N+1}(x[N] - \hat{A}[N - 1])^2 \quad (288)$$

위 식에서 보다시피 새로운 데이터 $x[N]$ 이 들어올 때마다 에러의 크기는 증가하는 것을 알 수 있다.

8.4.2 WLS → Sequential LS

다음으로 sequential LS 방법을 WLS(weighted LS)에 적용했을 때 어떻게 되는지 알아보자. 가중치 행렬 \mathbf{W} 가 만약 대각행렬로 주어져 있으며 각각의 성분이 white noise를 의미한다면 $[\mathbf{W}]_{ii} = \frac{1}{\sigma_i^2}$ 과 같이 쓸 수 있고 (??)에 의해 다음과 같다.

$$\hat{A}[N - 1] = \frac{\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}} \quad (289)$$

새로운 데이터 $x[N]$ 이 들어 왔을 때 LSE $\hat{A}[N]$ 은 다음과 같다.

$$\begin{aligned} \hat{A}[N] &= \frac{\sum_{n=0}^N \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} \\ &= \frac{\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2} + \frac{x[N]}{\sigma_N^2}}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} \\ &= \frac{\left(\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} \right) \hat{A}[N - 1]}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} + \frac{\frac{x[N]}{\sigma_N^2}}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} \\ &= \hat{A}[N - 1] - \frac{\frac{1}{\sigma_N^2} \hat{A}[N - 1]}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} + \frac{\frac{x[N]}{\sigma_N^2}}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} \end{aligned} \quad (290)$$

정리하면 LSE \hat{A} 에 대한 재귀식이 나온다.

$$\hat{A}[N] = \hat{A}[N-1] - \underbrace{\frac{\frac{1}{\sigma_N^2}}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} (x[N] - \hat{A}[N-1])}_{\text{correction}} \quad (291)$$

만약 모든 n 에 대하여 $\sigma_n^2 = \sigma$ 인 경우 (284)과 동일한 식이 얻어진다. **Correction 항은 새로운 데이터 $x[N]$ 의 불확실성에 의존한다.** 만약 새로운 데이터가 noisy하여 $\sigma_N \rightarrow \infty$ 인 경우 우리는 이전의 LSE 값을 업데이트하지 않는다. 반대로 새로운 데이터가 noise-free한 경우 $\sigma_N \rightarrow 0$ 이 되어 $\hat{A}[N] = x[N]$ 이 된다.

위 식을 조금 더 해석해보면 $x[n] = A + w[n]$ 이고 $w[n] \sim \mathcal{N}(0, \sigma^2)$ 인 데이터가 주어진 경우 WLS에 대한 sequential LSE \hat{A} 는 BLUE와 동일한 것을 알 수 있다. 따라서 (200)과 같이 다음 식이 성립한다.

$$\text{var}(\hat{A}[N-1]) = \frac{1}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}} \quad (292)$$

(291)에서 gain factor를 $K[N] = \frac{\frac{1}{\sigma_N^2}}{\sum_{n=0}^N \frac{1}{\sigma_n^2}}$ 와 같이 정의 한 후 전개하면 다음과 같다.

$$\begin{aligned} K[N] &= \frac{\frac{1}{\sigma_N^2}}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} \\ &= \frac{\frac{1}{\sigma_N^2}}{\frac{1}{\sigma_N^2} + \frac{1}{\text{var}(\hat{A}[N-1])}} \\ &= \frac{\text{var}(\hat{A}[N-1])}{\text{var}(\hat{A}[N-1]) + \sigma_N^2} \end{aligned} \quad (293)$$

Gain factor는 $0 \leq K[N] \leq 1$ 의 범위를 만족하며 $K[N]$ 또는 $\text{var}(\hat{A}[N-1])$ 값이 크다면 correction 값 또한 커진다. 그리고 만약 이전 $N-1$ 추정값의 분산이 작다면 correction 또한 작아진다.

분산 또한 다음과 같이 재귀적으로 나타낼 수 있다.

$$\begin{aligned} \text{var}(\hat{A}[N]) &= \frac{1}{\sum_{n=0}^N \frac{1}{\sigma_n^2}} \\ &= \frac{1}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} + \frac{1}{\sigma_N^2}} \\ &= \frac{1}{\frac{1}{\text{var}(\hat{A}[N-1])} + \frac{1}{\sigma_N^2}} \\ &= \frac{\text{var}(\hat{A}[N-1])\sigma_N^2}{\text{var}(\hat{A}[N-1]) + \sigma_N^2} \\ &= \left(1 - \frac{\text{var}(\hat{A}[N-1])}{\text{var}(\hat{A}[N-1]) + \sigma_N^2}\right) \text{var}(\hat{A}[N-1]) \end{aligned} \quad (294)$$

정리하면 분산에 대한 재귀식이 나온다.

$$\text{var}(\hat{A}[N]) = (1 - K[N])\text{var}(\hat{A}[N-1]) \quad (295)$$

지금까지 구한 재귀식들을 모아보면 한 번에 쓰면 다음과 같다.

Initial Value:

$$\hat{A}[0] = x[0]$$

$$\text{var}(\hat{A}[0]) = \sigma_0^2$$

Estimator Update:

$$\hat{A}[N] = \hat{A}[N-1] - K[N](x[N] - \hat{A}[N-1]) \quad (296)$$

where,

$$K[N] = \frac{\text{var}(\hat{A}[N-1])}{\text{var}(\hat{A}[N-1]) + \sigma_N^2}$$

Variance Update:

$$\text{var}(\hat{A}[N]) = (1 - K[N])\text{var}(\hat{A}[N-1])$$

8.4.3 WLS → Sequential LS (vector parameter)

다음으로 sequential LS를 벡터 파라미터에 적용해보자. 이전 섹션과 동일하게 가중치 행렬 \mathbf{W} 가 대각행렬이면서 각각의 성분이 white noise의 역수 $1/\sigma_i^2$ 을 의미한다면 $[\mathbf{W}]_{ii} = 1/\sigma_i^2$ 과 같이 나타낼 수 있다. 가중치 행렬의 역행렬은 $\mathbf{W} = \mathbf{C}^{-1}$ 와 같고 \mathbf{C} 은 공분산 행렬이 된다. 벡터 파라미터에 대한 WLS의 LS criterion J 은 다음과 같다.

$$J = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^\top \mathbf{C}^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \quad (297)$$

이러한 가정은 BLUE의 (??)와 동일하다. (??)에서 이미 WLS에 대한 LSE를 구하였다.

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{C}^{-1} \mathbf{x} \quad (298)$$

공분산 $\mathbf{C}_{\hat{\boldsymbol{\theta}}}$ 은 다음과 같다.

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^\top \mathbf{C}^{-1} \mathbf{H})^{-1} \quad (299)$$

만약 \mathbf{C} 가 대각행렬이 아닌 경우(즉, white noise가 아닌 경우) 위 식들은 재귀적으로 계산되지 않는다. 따라서 \mathbf{C} 가 대각행렬이라는 가정 하에 n 번째 데이터가 들어온 경우 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \mathbf{C}[n] &= \text{diag}(\sigma_0^2, \sigma_1^2, \dots, \sigma_n^2) \\ \mathbf{H}[n] &= \begin{bmatrix} \mathbf{H}[n-1] \\ \mathbf{h}^\top[n] \end{bmatrix} = \begin{bmatrix} n \times p \\ 1 \times p \end{bmatrix} \\ \mathbf{x}[n] &= [x[0], x[1], \dots, x[n]]^\top \end{aligned} \quad (300)$$

(298), (299)도 n 번째 데이터에 대한 형태로 나타내면 다음과 같다.

$$\hat{\boldsymbol{\theta}}[n] = (\mathbf{H}^\top[n] \mathbf{C}^{-1}[n] \mathbf{H}[n])^{-1} \mathbf{H}^\top[n] \mathbf{C}^{-1}[n] \mathbf{x}[n] \quad (301)$$

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \Sigma[n] = (\mathbf{H}^\top[n] \mathbf{C}^{-1}[n] \mathbf{H}[n])^{-1} \quad (302)$$

최종적인 벡터 파라미터의 sequential LSE는 다음과 같다.

Estimator Update:

$$\hat{\boldsymbol{\theta}}[n] = \hat{\boldsymbol{\theta}}[n-1] + \mathbf{K}[n](x[n] - \mathbf{h}^\top[n]\hat{\boldsymbol{\theta}}[n-1])$$

where,

$$\mathbf{K}[n] = \frac{\Sigma[n-1]\mathbf{h}[n]}{\sigma_n^2 + \mathbf{h}^\top[n]\Sigma[n-1]\mathbf{h}[n]} \quad (303)$$

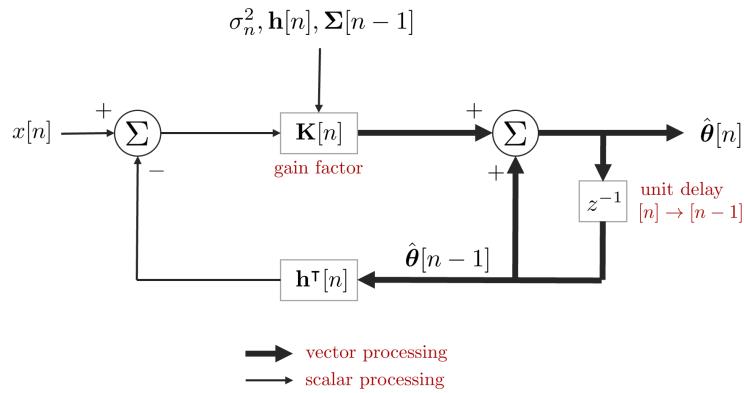
Variance Update:

$$\Sigma[n] = (\mathbf{I} - \mathbf{K}[n])\Sigma[n-1]$$

- $\mathbf{K}[n] \in \mathbb{R}^{p \times 1}$

- $\Sigma \in \mathbb{R}^{p \times p}$

위 식에서 어떠한 역행렬도 구하지 않는다는 점을 주목하자. Sequential LS에 대한 파이프라인을 그리면 아래 그림과 같다.



Sequential LS가 위와 같이 재귀적으로 돌면서 ($\mathbf{H}^\top[n-1]\mathbf{C}^{-1}[n-1]\mathbf{H}[n-1]$)은 (301)에 의해 반드시 역행렬이 존재해야 한다. 이는 $\mathbf{H}[n-1]$ 의 역행렬이 존재해야 하는 것과 동치이며 $n \times p$ 크기의 행렬인 \mathbf{H} 의 rank는 p 보다 커야한다는 것을 의미한다. 결론적으로 $n > p$ 를 만족해야 하므로 데이터의 개수 n 이 파라미터의 개수 p 보다 많은 over-determined system이 되어야 한다.

8.5 Constrained Least Squares

이번 섹션에서는 LS 문제 중 미지의 파라미터가 제약조건이 있는 경우에 대하여 배운다. 예를 들어, 특정 신호의 크기를 추정하고자 하는데 몇몇 크기가 동일한 것을 미리 알고 있다는 제약조건을 추가한 것과 동일하다. **만약 제약조건이 선형(linear)인 경우 상대적으로 쉽게 해결할 수 있다.** 선형인 제약조건 케이스에 대하여 알아보자.

LS criterion이 다음과 같이 주어졌다고 가정하자.

$$J_c = (\mathbf{x} - \mathbf{H}\theta)^\top(\mathbf{x} - \mathbf{H}\theta) \quad (304)$$

그리고 $\mathbf{A}\theta = \mathbf{b}$ 와 같은 선형(linear) 제약조건이 주어졌다고 하자. $\mathbf{A} \in \mathbb{R}^{r \times p}$ 이고 $\theta \in \mathbb{R}^{p \times 1}$ 이며 $\mathbf{b} \in \mathbb{R}^{r \times 1}$ 이다. Lagrangian multiplier λ 를 곱하여 LS criterion을 다시 쓰면 다음과 같다.

$$J_c = (\mathbf{x} - \mathbf{H}\theta)^\top(\mathbf{x} - \mathbf{H}\theta) + \lambda^\top(\mathbf{A}\theta - \mathbf{b}) \quad (305)$$

- $\lambda \in \mathbb{R}^{r \times 1}$: Lagrangian multiplier

위 식을 전개하면 다음과 같다.

$$J_c = \mathbf{x}^\top \mathbf{x} - 2\theta^\top \mathbf{H}^\top \mathbf{x} + \theta^\top \mathbf{H}^\top \mathbf{H}\theta + \lambda^\top \mathbf{A}\theta - \lambda^\top \mathbf{b} \quad (306)$$

찾고자 하는 파라미터 θ 에 대하여 편미분을 수행하면 다음과 같다.

$$\frac{\partial J_c}{\partial \theta} = -2\mathbf{H}^\top \mathbf{x} + 2\mathbf{H}^\top \mathbf{H}\theta + \mathbf{A}^\top \lambda \quad (307)$$

위 식을 0으로 놓고 제약조건에 대한 LSE $\hat{\theta}_c$ 를 구하면 다음과 같다.

$$\begin{aligned} \hat{\theta}_c &= (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x} - \frac{1}{2} (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top \lambda \\ &= \hat{\theta} - (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top \frac{\lambda}{2} \end{aligned} \quad (308)$$

$\hat{\theta}$ 는 제약조건이 없는 경우(unconstrained) LSE를 의미한다. 적절한 λ 값을 찾기 위해 좌항에 \mathbf{A} 를 곱하면 다음과 같다.

$$\mathbf{A}\hat{\theta}_c = \mathbf{A}\hat{\theta} - \mathbf{A}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top \frac{\lambda}{2} = \mathbf{b} \quad (309)$$

λ 에 대해 정리하면 다음과 같다.

$$\frac{\lambda}{2} = [\mathbf{A}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top]^{-1} (\mathbf{A}\hat{\theta} - \mathbf{b}) \quad (310)$$

위 식을 (308)에 대입하면 다음과 같은 식을 얻을 수 있다.

$$\boxed{\hat{\theta}_c = \hat{\theta} - (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top [\mathbf{A}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{A}^\top]^{-1} (\mathbf{A}\hat{\theta} - \mathbf{b})} \quad (311)$$

- $\hat{\theta} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x}$

제약조건이 있는 LSE $\hat{\theta}_c$ 는 제약조건이 없는 LSE $\hat{\theta}$ 에서 약간 수정된 버전과 같다. 운이 좋게도 $\hat{\theta}$ 가 $\mathbf{A}\hat{\theta} = \mathbf{b}$ 의 제약조건을 만족하는 경우 $\hat{\theta}_c = \hat{\theta}$ 가 된다. 물론 이러한 경우는 잘 발생하지 않는다.

8.6 Nonlinear Least Squares

이번 섹션에서는 비선형 LS 문제에 대해 배운다. 비선형 신호 모델 $\mathbf{s}(\boldsymbol{\theta})$ 에 대한 LS criterion은 다음과 같다.

$$J = (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^\top (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) \quad (312)$$

위 식에서 만약 $\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}) \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ 를 만족하면 LSE는 MLE와 동일해진다. 일반적으로 비선형 함수가 주어진 경우 LS 문제는 매우 풀기 어려워지거나 거의 불가능하다. 위와 같이 비선형의 J 를 최소화 문제는 비선형 회귀 방법(nonlinear regression problem)으로도 잘 알려져 있으며 수많은 이론적 연구가 존재한다 [Bard 1974, Seber and Wild 1989]. 현실적으로 이를 풀기 위해서는 반드시 반복적인 방법(Iterative method)을 사용하여 풀어야 하며 챕터 7 MLE의 수치적인 해를 구하는 것과 동일한 한계점이 존재한다(=수렴이 안되는 경우가 생긴다).

일반적인 비선형 LSE의 해법을 말하기 전에 문제의 복잡도를 줄이는 두 가지 방법론에 대해 설명한다.

- Method 1: 파라미터 변환 (transformation of parameters)
- Method 2 : 파라미터 분리 (separability of parameters)

Method 1: transformation of paramters

첫 번째 방법은 파라미터 $\boldsymbol{\theta}$ 를 일대일 변환하여 선형 신호 모델로 변경하는 방법이다.

$$\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta}) \quad (313)$$

함수 \mathbf{g} 는 $\boldsymbol{\theta}$ 에 대해 p 차원의 크기를 가진 벡터 함수이며 역함수가 존재해야 한다. 만약 \mathbf{g} 를 찾을 수 있다면 다음과 같이 $\boldsymbol{\alpha}$ 에 대한 선형 모델로 변환할 수 있게 된다.

$$\mathbf{s}(\boldsymbol{\theta}(\boldsymbol{\alpha})) = \mathbf{s}(\mathbf{g}^{-1}(\boldsymbol{\alpha})) = \mathbf{H}\boldsymbol{\alpha} \quad (314)$$

그 다음부터는 $\boldsymbol{\alpha}$ 에 대한 선형 LSE $\hat{\boldsymbol{\alpha}}$ 를 정규 방정식을 통해 쉽게 구할 수 있다. 최종적으로 비선형 LSE $\hat{\boldsymbol{\theta}}$ 는 다음과 같이 구한다.

$$\hat{\boldsymbol{\theta}} = \mathbf{g}^{-1}(\hat{\boldsymbol{\alpha}}) \quad (315)$$

- $\hat{\boldsymbol{\alpha}} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x}$

위와 같은 방법론은 다른 변환 공간에서 최적값을 1차로 구하고 일대일 매핑 후 2차로 최적값을 찾는 방법에 의존한다. 일반적으로 \mathbf{g} 함수를 찾는 것은 매우 어려우며 소수의 LS 문제만 이를 찾을 수 있다.

Method 2 : separability of parameters

두 번째 방법은 비선형 함수 \mathbf{s} 를 선형인 성분과 비선형인 성분으로 분리하는 방법을 말한다.

$$\mathbf{s} = \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta} \quad (316)$$

- $\mathbf{H}(\boldsymbol{\alpha}) \in \mathbb{R}^{N \times q}$

이 때 추정하고자 하는 파라미터는 $\boldsymbol{\theta}$ 는 다음과 같다.

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} (p-q) \times 1 \\ q \times 1 \end{bmatrix} \quad (317)$$

따라서 신호 모델은 $\boldsymbol{\alpha}$ 는 비선형인데 $\boldsymbol{\beta}$ 는 선형인데 된다. LS criterion J 는 $\boldsymbol{\beta}$ 에 대하여 1차적으로 최소화된 다음 $\boldsymbol{\alpha}$ 에 대하여 2차적으로 최소화한다.

$$J(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{x} - \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta})^\top (\mathbf{x} - \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta}) \quad (318)$$

$\boldsymbol{\beta}$ 에 대해 J 를 먼저 최소화하면 다음과 같은 정규 방정식의 해를 얻는다.

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^\top(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha}))^{-1} \mathbf{H}^\top(\boldsymbol{\alpha})\mathbf{x} \quad (319)$$

그리고 $\hat{\beta}$ 을 대입하여 (280)의 다섯번째 줄과 동일한 형태로 표현하면 다음과 같다.

$$J(\alpha, \hat{\beta}) = \mathbf{x}^\top [\mathbf{I} - \mathbf{H}(\alpha)(\mathbf{H}^\top(\alpha)\mathbf{H}(\alpha))^{-1}\mathbf{H}^\top(\alpha)]\mathbf{x} \quad (320)$$

결론적으로 비선형 LS 문제는 아래 식을 최대화하는 문제로 변형된다.

$$J(\alpha, \hat{\beta}) \propto \arg \max_{\alpha} \left[\mathbf{H}(\alpha)(\mathbf{H}^\top(\alpha)\mathbf{H}(\alpha))^{-1}\mathbf{H}^\top(\alpha)\mathbf{x} \right] \quad (321)$$

8.6.1 General approach for nonlinear LS

이번 섹션에서는 앞서 설명한 두 가지 트릭이 적용되지 않는 일반적인 비선형 LS 문제를 다룬다. 우선 비선형 LS criterion을 다시 써보면 다음과 같다.

$$J = (\mathbf{x} - \mathbf{s}(\theta))^\top (\mathbf{x} - \mathbf{s}(\theta)) \quad (322)$$

위 식을 미분해보자.

$$\frac{\partial J}{\partial \theta_j} = -2 \sum_{i=0}^{N-1} (x[i] - s[i]) \frac{\partial s[i]}{\partial \theta_j} = 0 \quad (323)$$

- $j = 1, 2, \dots, p$

그리고 $N \times p$ 크기의 자코비언 행렬을 다음과 같이 정의한다.

$$\left[\frac{\partial \mathbf{s}(\theta)}{\partial \theta} \right]_{ij} = \frac{\partial s[i]}{\partial \theta_j} \quad (324)$$

- $i = 0, 1, \dots, N-1$

- $j = 1, 2, \dots, p$

(323) 식에 대입하여 0이 되도록 설정하면 다음과 같다.

$$\begin{aligned} \sum_{i=0}^{N-1} (x[i] - s[i]) \left[\frac{\partial s(\theta)}{\partial \theta} \right]_{ij} &= 0 \\ \frac{\partial \mathbf{s}(\theta)^\top}{\partial \theta} (\mathbf{x} - \mathbf{s}(\theta)) &= 0 \quad \dots \text{in matrix form} \end{aligned} \quad (325)$$

위 식은 p 차원 크기를 가진 비선형 방정식들의 집합으로 볼 수 있다.

8.6.2 Newton-Raphson iteration

위 식을 $\mathbf{g}(\theta)$ 라고 하자.

$$\mathbf{g}(\theta) = \frac{\partial \mathbf{s}(\theta)^\top}{\partial \theta} (\mathbf{x} - \mathbf{s}(\theta)) \quad (326)$$

위 식에 Newton-Raphson 방법을 적용하면 다음과 같다.

$$\theta_{k+1} = \theta_k - \left(\frac{\partial \mathbf{g}(\theta)}{\partial \theta} \right)^{-1} \mathbf{g}(\theta) \Big|_{\theta=\theta_k} \quad (327)$$

위 식에서 $\mathbf{g}(\theta)$ 에 대한 자코비언 행렬을 찾아야 하는데 이는 정의 상 J 의 해시안(hessian) 행렬을 구하는 것과 동일하다.

$$\frac{\partial [\mathbf{g}(\theta)]_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \left[\sum_{n=0}^{N-1} (x[n] - s[n]) \frac{\partial s[n]}{\partial \theta_i} \right] \quad (328)$$

편미분을 전개하면 다음과 같다.

$$\frac{\partial [\mathbf{g}(\theta)]_i}{\partial \theta_j} = \sum_{n=0}^{N-1} \left[(x[n] - s[n]) \frac{\partial^2 s[n]}{\partial \theta_i \partial \theta_j} - \frac{\partial s[n]}{\partial \theta_i} \frac{\partial s[n]}{\partial \theta_j} \right] \quad (329)$$

보다 깔끔한 수식 표현을 위해 다음과 같이 치환한다.

$$[\mathbf{H}(\boldsymbol{\theta})]_{ij} = \left[\frac{\partial s(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]_{ij} = \frac{\partial s[i]}{\partial \theta_j} \quad (330)$$

$$[\mathbf{G}_n(\boldsymbol{\theta})]_{ij} = \frac{\partial^2 s[n]}{\partial \theta_i \partial \theta_j} \quad (331)$$

치환된 문자로 전개하면 다음과 같다.

$$\begin{aligned} \frac{\partial [\mathbf{g}(\boldsymbol{\theta})]_i}{\partial \theta_j} &= \sum_{n=0}^{N-1} (x[n] - s[n]) [\mathbf{G}_n(\boldsymbol{\theta})]_{ij} - [\mathbf{H}(\boldsymbol{\theta})]_{nj} [\mathbf{H}(\boldsymbol{\theta})]_{ni} \\ &= \sum_{n=0}^{N-1} [\mathbf{G}_n(\boldsymbol{\theta})]_{ij} (x[n] - s[n]) - [\mathbf{H}^\top(\boldsymbol{\theta})]_{in} [\mathbf{H}(\boldsymbol{\theta})]_{nj} \end{aligned} \quad (332)$$

다음식과 같이 벡터 형태로 표현한다.

$$\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{n=0}^{N-1} (x[n] - s[n]) \mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{H}^\top(\boldsymbol{\theta}) \mathbf{H}(\boldsymbol{\theta}) \quad (333)$$

최종적으로 (327)에 위 식을 대입하면 Newton-Rapshon 수식이 완성된다.

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \left(\mathbf{H}^\top(\boldsymbol{\theta}) \mathbf{H}(\boldsymbol{\theta}) - \sum_{n=0}^{N-1} (x[n] - s[n]) \mathbf{G}_n(\boldsymbol{\theta}) \right)^{-1} \cdot \mathbf{H}^\top(\boldsymbol{\theta}_k) (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}_k))} \quad (334)$$

\mathbf{H}, \mathbf{G} 행렬은 각각 선호 함수 s 를 파라미터 $\boldsymbol{\theta}$ 에 대하여 1차, 2차 미분을 수행한 행렬이다(=jacobian, hessian). 만약 선호 모델이 선형이어서 $\mathbf{s}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}$ 를 만족한다면 $\mathbf{G}(\boldsymbol{\theta}) = 0$ 이 되고 $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{H}$ 가 되어 위 식은 다음과 같이 변한다.

$$\begin{aligned} \boldsymbol{\theta}_{k+1} &= \boldsymbol{\theta}_k + (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}_k) \\ &= (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{x} \end{aligned} \quad (335)$$

위와 같이 문제가 선형이라면 반복적(iterative)으로 풀지 않아도 한 번에 close form solution이 나오게 된다. 비록 선형이 아니더라도 선형에 근사한 경우 위 반복식은 빠르게 수렴한다.

8.6.3 Gauss-Newton iteration

또 다른 방법은 비선형 LS 문제를 선형으로 근사한 후 선형 LS 문제를 푸는 방법이 있다. Newton-Raphson 방법과 차이점은 J 의 미분값이 현재 상태에 대하여 선형화(linearized)된다는 것이다. 보다 정확한 차이점 이해를 위해 스칼라 파리미터에 대한 예시를 들어보자. 현재 상태를 θ_0 라고 했을 때 $s[n; \theta]$ 는 다음과 같이 선형화된다.

$$s[n; \theta] \approx s[n; \theta_0] + \frac{\partial s[n; \theta]}{\partial \theta} \Big|_{\theta=\theta_0} (\theta - \theta_0) \quad (336)$$

- $s[n; \theta]$: 스칼라 함수 $s(\theta)$ 의 n 번째 값

비선형 LS criterion은 다음과 같이 근사화된다.

$$\begin{aligned} J &= \sum_{n=0}^{N-1} (x[n] - s[n; \theta])^2 \\ &\approx \sum_{n=0}^{N-1} \left(x[n] - s[n; \theta_0] + \frac{\partial s[n; \theta]}{\partial \theta} \Big|_{\theta=\theta_0} \theta_0 - \frac{\partial s[n; \theta]}{\partial \theta} \Big|_{\theta=\theta_0} \theta \right)^2 \\ &= (\mathbf{x} - \mathbf{s}(\theta_0) + \mathbf{H}(\theta_0)\theta_0 - \mathbf{H}(\theta_0)\theta)^\top (\mathbf{x} - \mathbf{s}(\theta_0) + \mathbf{H}(\theta_0)\theta_0 - \mathbf{H}(\theta_0)\theta) \end{aligned} \quad (337)$$

$\mathbf{x} - \mathbf{s}(\theta_0) + \mathbf{H}(\theta_0)\theta_0$ 값은 이미 알고 있는 값이므로 LSE $\hat{\theta}$ 는 다음과 같이 구할 수 있다.

$$\begin{aligned} \hat{\theta} &= (\mathbf{H}^\top(\theta_0) \mathbf{H}(\theta_0))^{-1} \mathbf{H}^\top(\theta_0) (\mathbf{x} - \mathbf{s}(\theta_0) + \mathbf{H}(\theta_0)\theta_0) \\ &= \theta_0 + (\mathbf{H}^\top(\theta_0) \mathbf{H}(\theta_0))^{-1} \mathbf{H}^\top(\theta_0) (\mathbf{x} - \mathbf{s}(\theta_0)) \end{aligned} \quad (338)$$

위 식을 반복(iteration)식으로 변경하면 이는 다음과 같다.

$$\hat{\theta}_{k+1} = \theta_k + (\mathbf{H}^\top(\theta_k)\mathbf{H}(\theta_k))^{-1}\mathbf{H}^\top(\theta_k)(\mathbf{x} - \mathbf{s}(\theta_k)) \quad (339)$$

위 식은 Newton-Rapshon 법과 유사하지만 2차 미분된 헤시안 행렬 \mathbf{G}_n 이 존재하지 않는다. 이러한 선형화 방법을 Gauss-Newton 방법이라고 하며 일반적인 벡터 파라미터에 대한 형태로 표현하면 다음과 같다.

$$\boxed{\theta_{k+1} = \theta_k + (\mathbf{H}^\top(\theta_k)\mathbf{H}(\theta_k))^{-1}\mathbf{H}^\top(\theta_k)(\mathbf{x} - \mathbf{s}(\theta_k))} \quad (340)$$

$$- \left[\mathbf{H}(\theta_k) \right]_{ij} = \frac{\partial s[i]}{\partial \theta_j}$$

9 Method of Moments

TBD...

10 The Bayesian Philosophy

본 챕터에서는 지금까지 설명한 고전적인 추정 방법에서 벗어나 파라미터 θ 또한 하나의 확률 변수(random variable)로 보는 베이지안(Bayesian) 관점에 대하여 설명한다. 앞서 챕터 1에서 설명하였듯이 고전적인 관점과 베이지안 관점의 차이는 다음과 같다.

$$\begin{aligned} \text{Frequentist: } & \underbrace{x[n]}_{\text{r.v.}} = \underbrace{\theta}_{\text{deterministic}} + w[n] \\ \text{Bayesian: } & \underbrace{x[n]}_{\text{r.v.}} = \underbrace{\theta}_{\text{r.v.}} + w[n] \end{aligned} \quad (341)$$

- 고전적인 추정 방법 : 파라미터 θ 를 미지의 결정론적(deterministic) 파라미터로 보는 빈도주의(frequentist) 관점으로 해석
- 현대적인 추정 방법 : 파라미터 θ 를 별도의 확률 변수(random variable)로 보는하는 베이지안(bayesian) 관점으로 해석

베이지안 철학의 모티브는 다음과 같다.

1. 만약 우리가 θ 에 대한 사전(prior) 정보를 알고 있다면 이는 더 나은 추정에 활용될 수 있다. 하지만 이를 위해서는 θ 에 대한 prior pdf가 미리 주어져 있거나 계산할 수 있어야 한다.
2. Bayesian 추정은 MVUE를 찾는 것이 불가능한 경우 사용하기 좋은 방법이다. 예를 들어 특정 불편추정값(unbiased estimator)의 분산이 다른 불편추정값들의 분산보다 일관되게 작지 않은 경우를 생각해보자. 이런 경우 고전적인 방법으로는 MVUE를 찾는 것이 불가능하므로 파라미터 θ 에 pdf를 적용함으로써 우리는 그 추정값 $\hat{\theta}$ 을 찾는 방법을 생각할 수 있다. $\hat{\theta}$ 가 다른 추정값들보다 평균제곱오차(MSE)가 작다면 이는 최적의 추정값이라고 결론지을 수 있다. 즉, 파라미터 θ 에 대한 확률적인 가정을 함으로써 차선의 추정값을 찾을 수 있는데 이러한 관점이 Bayesian 관점이다.

10.1 Prior Knowledge and Estimation

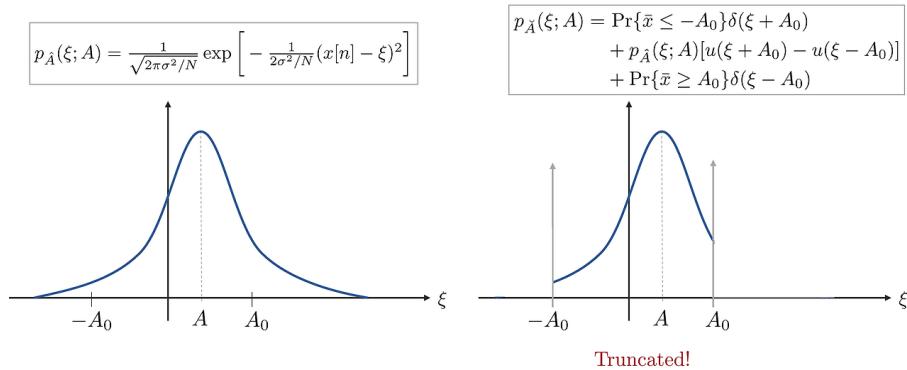
만약 파라미터에 대한 사전 지식(prior knowledge)이 주어졌다면 이는 더 정확한 추정값을 찾는데 활용할 수 있다. Example 3.1에서 파라미터 A 에 대한 MVUE는 \bar{x} 라고 하였다. 하지만 이는 파라미터 A 를 $-\infty < A < \infty$ 에서 얻은 경우에 대해서만 참이다. 만약 A 를 제한된 영역인 $-A_0 \leq A \leq A_0$ 에서 얻었다고 가정해보자. 이렇게 되면 정해진 영역 밖에서는 $\hat{A} = \bar{x}$ 로 보는 것이 부적절할 수 있다. 다음과 같은 잘린 영역의(truncated) 추정값을 사용하면 추정값의 성능을 향상시킬 수 있다.

$$\check{A} = \begin{cases} -A_0 & \cdots \bar{x} < -A_0 \\ \bar{x} & \cdots -A_0 \leq \bar{x} \leq A_0 \\ A_0 & \cdots \bar{x} > A_0 \end{cases} \quad (342)$$

위 추정값에 대한 pdf를 구해보면 다음과 같다.

$$\begin{aligned} p_A(\xi; A) &= \Pr\{\bar{x} \leq -A_0\}\delta(\xi + A_0) \\ &\quad + p_{\check{A}}(\xi; A)[u(\xi + A_0) - u(\xi - A_0)] \\ &\quad + \Pr\{\bar{x} \geq A_0\}\delta(\xi - A_0) \end{aligned} \quad (343)$$

- $u(x)$: 단위 스텝(unit step) 함수



위 그림에서 보다시피 \check{A} 추정값은 편향되어 있다(biased). 두 추정값을 MSE를 통해 비교해보면 다음과 같다.

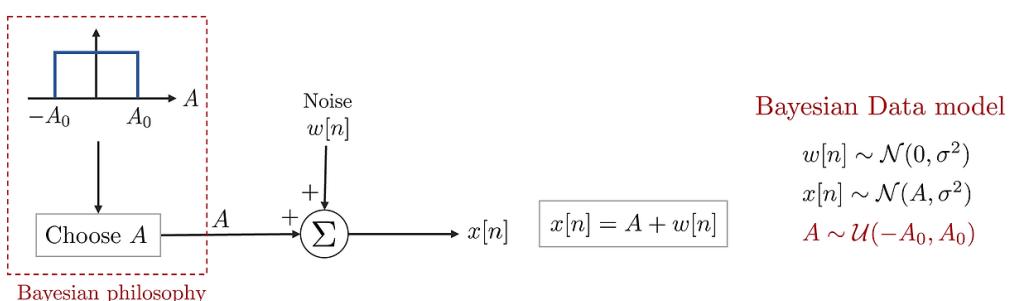
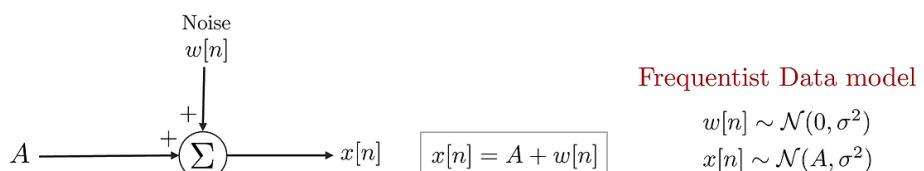
$$\begin{aligned}
 \text{mse}(\hat{A}) &= \int_{-\infty}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi; A) d\xi \\
 &= \int_{-\infty}^{-A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A) d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A) d\xi \\
 &\quad + \int_{A_0}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi; A) d\xi \\
 &> \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A) d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A) d\xi \\
 &\quad + \int_{A_0}^{\infty} (A_0 - A)^2 p_{\hat{A}}(\xi; A) d\xi \\
 &= \text{mse}(\check{A})
 \end{aligned} \tag{344}$$

$$\text{mse}(\hat{A}) > \text{mse}(\check{A}) \tag{345}$$

따라서 truncated 추정값 \check{A} 가 평균 추정값 \hat{A} 보다 MSE 측면에서 더 좋은 값인 것을 알 수 있다. 비록 \hat{A} 는 여전히 MVUE이지만, 여기에 편향성(biased)이 추가된 \check{A} 를 사용하여 MSE를 더 줄일 수 있었다.

우리는 방금 \check{A} 와 같은 좋은 추정값을 얻었지만 몇몇 독자들은 여전히 \check{A} 보다 정확한 최적의 추정값이 존재하는지 궁금할 수 있다. 베이지안 추정 방법으로 데이터 모델을 재구성해야만 이를 확인할 수 있다. A 가 정해진 영역에서 부터 얻은 값이라는 것을 알기 때문에 파라미터 A 을 해당 영역에서 랜덤으로 추출한 확률 변수(random variable)로 가정할 수 있다. 특정 분포를 설정한 만한 가정이 없으므로 균일 분포(uniform distribution)이라고 가정하면 $A \sim U[-A_0, A_0]$ 이 된다.

기준의 고전적인 추정 방법과 베이지안 추정 방법을 비교한 그림은 아래와 같다.



두 관점 모두 궁극적으로는 파라미터 A 에 대한 최적의 추정값 \hat{A} 을 얻는 것이 목표이지만 베이지안 추정 관점에서 봤을 때 A 에 대한 사전 정보(prior knowledge)를 확인할 수 있게 되었다. 예를 들어 우리는 Bayesian MSE(=BMSE)를 최소화시키는 추정값 \hat{A} 를 찾을 수 있다. BMSE는 다음과 같이 정의된다.

$$\boxed{\text{Bmse}(\hat{A}) = \mathbb{E}[(A - \hat{A})^2]} \quad (346)$$

기존 MSE의 에러 $\hat{A} - A$ 와는 달리 BMSE에서는 에러를 $A - \hat{A}$ 로 정의하였다. 이러한 정의는 추후 베이지안 추정값에 대한 벡터 공간을 해석 할 때 유용하게 사용된다. 위 식에서 A 는 고정된 값이 아닌 확률 변수(random variable)임에 유의하자. 따라서 기대값은 joint pdf $p(\mathbf{x}, A)$ 에 대한 기대값이 된다. 이는 근본적으로 기존 MSE와는 다른 관점이다.

$$\begin{aligned} \text{mse}(\hat{A}) &= \int (\hat{A} - A)^2 p(\mathbf{x}; A) d\mathbf{x} \\ \text{Bmse}(\hat{A}) &= \int \int (A - \hat{A})^2 p(\mathbf{x}, A) d\mathbf{x} dA \end{aligned} \quad (347)$$

joint pdf는 $p(\mathbf{x}, A) = p(A|\mathbf{x})p(\mathbf{x})$ 와 같이 분리할 수 있으므로 위 식을 다시 쓰면 다음과 같다.

$$\text{Bmse}(\hat{A}) = \int \left[\int (A - \hat{A})^2 p(A|\mathbf{x}) dA \right] p(\mathbf{x}) d\mathbf{x} \quad (348)$$

확률의 정의에 의해 모든 \mathbf{x} 에 대하여 $p(\mathbf{x}) > 0$ 이 성립한다. 따라서 위 브라켓 안에 있는 값만 작아진다면 BMSE는 최소화될 수 있다. 따라서 브라켓 안에 있는 식을 미분하면 다음과 같다.

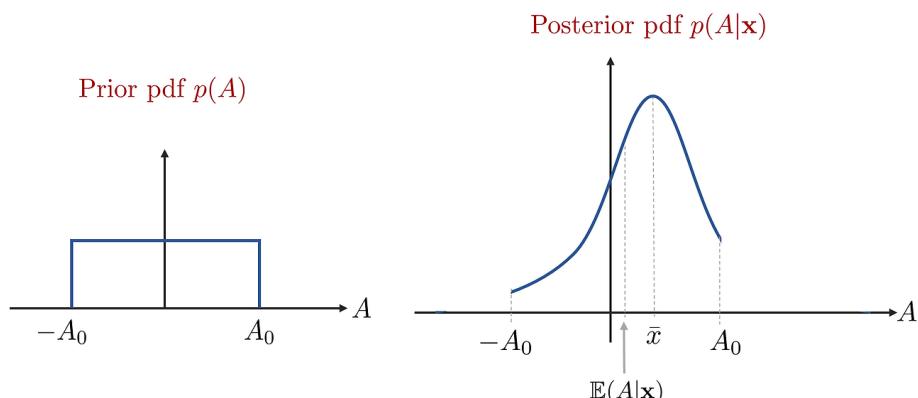
$$\begin{aligned} \frac{\partial}{\partial \hat{A}} \int (A - \hat{A})^2 p(A|\mathbf{x}) dA &= \int \frac{\partial}{\partial \hat{A}} (A - \hat{A})^2 p(A|\mathbf{x}) dA \\ &= \int -2(A - \hat{A}) p(A|\mathbf{x}) dA \\ &= -2 \int A p(A|\mathbf{x}) dA + 2\hat{A} \underbrace{\int p(A|\mathbf{x}) dA}_{=1} \end{aligned} \quad (349)$$

위 식을 0으로 놓고 풀면 다음과 같은 추정값을 얻는다.

$$\hat{A} = \int A p(A|\mathbf{x}) dA \quad (350)$$

$$\therefore \hat{A} = \mathbb{E}(A|\mathbf{x}) \quad (351)$$

따라서 BMSE에 대한 최적의 추정값 \hat{A} 는 posterior pdf $p(A|\mathbf{x})$ 의 기대값인 것을 알 수 있으며 이를 MMSE(minimum MSE) 추정값이라고 부른다. Posterior pdf는 데이터 \mathbf{x} 가 관찰되었을 때 파라미터 A 의 pdf를 의미한다. 이와 대조적으로 prior pdf $p(A)$ 는 데이터가 관측되기 전 A 의 pdf를 의미한다. 직관적으로 이해하자면 데이터를 관측하는 행동이 A 의 pdf를 뾰족하게 만드는 효과를 가진다. 이는 데이터에 대한 지식이 A 의 불확실성을 줄여주기 때문이다.



MMSE 추정값을 구하는 작업은 간단한 작업이 아니다. 우선 posterior pdf를 bayes rule을 통해 분해하는 것부터 시작하자.

$$\begin{aligned}
p(A|x) &= \frac{p(x|A)p(A)}{p(x)} \\
&= \frac{p(x|A)p(A)}{\int p(x|A)p(A)dA} \tag{352}
\end{aligned}$$

위 식에서 분모는 파라미터 A 와 무관하여 전체 확률의 크기를 정의에 따라 1로 만들어주는 정규화(normalization) 값이다. $p(A)$ 값은 앞서 균일 분포 $\mathcal{U}(-A_0, A_0)$ 를 따른다고 하였다. $p(x|A)$ 는 $x[n] = A + w[n]$ 에서 x 의 pdf를 구하는 것과 동일하므로 아래와 같은 likelihood가 얻어진다.

$$p(x|\mathbf{A}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \tag{353}$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

위 식은 이전 챕터에서 설명한 pdf $p(x; A)$ 와 동일한 것을 알 수 있다. 고전적인 추정 방법으로 해석할 때는 $p(x; A)$ 을 사용하는 것이 맞지만 베이지안 추정 방법에서 해석할 때는 $p(x|A)$ 를 사용해야 한다. Prior와 likelihood를 (352)에 대입하면 다음 식이 얻어진다.

$$p(A|x) = \begin{cases} \frac{1}{c\sqrt{2\pi\frac{\sigma^2}{N}}} \exp \left[-\frac{1}{2\frac{\sigma^2}{N}}(A - \bar{x})^2 \right] & |A| \leq A_0 \\ 0 & |A| > A_0 \end{cases} \tag{354}$$

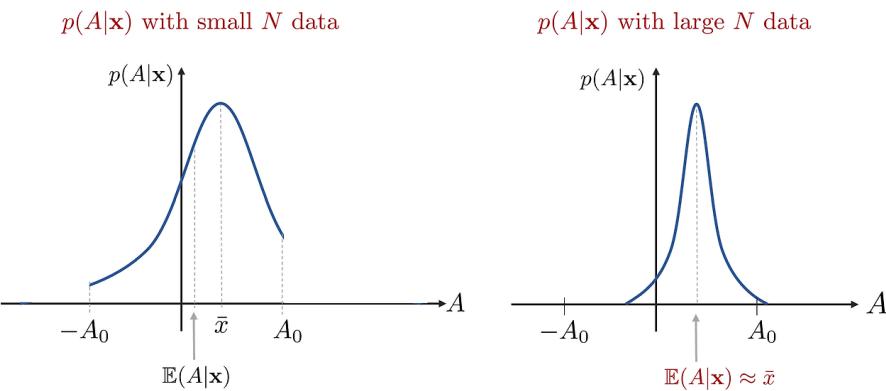
c 는 pdf의 크기를 1로 만들어주는 정규화 값이며 다음과 같다.

$$c = \int_{-A_0}^{A_0} \frac{1}{2\pi\frac{\sigma^2}{N}} \exp \left[-\frac{1}{2\frac{\sigma^2}{N}}(A - \bar{x})^2 \right] dA \tag{355}$$

따라서 MMSE \hat{A} 는 다음과 같이 구할 수 있다.

$$\begin{aligned}
\hat{A} &= \mathbb{E}(A|x) \\
&= \int_{-\infty}^{\infty} Ap(A|x)dA \\
&= \frac{1}{c} \int_{-A_0}^{A_0} A \frac{1}{2\pi\frac{\sigma^2}{N}} \exp \left[-\frac{1}{2\frac{\sigma^2}{N}}(A - \bar{x})^2 \right] dA
\end{aligned}
\tag{356}$$

MMSE 추정값 \hat{A} 은 비록 close form으로 구할 수는 없지만 적어도 \hat{A} 가 \bar{x}, σ^2, A_0 에 관한 함수라는 것을 알 수 있다. \hat{A} 값은 A 에 대한 truncated 사전 분포 $p(A)$ 가 반영되었기 때문에 평균 \bar{x} 의 값과 달리 편향되어 있다(biased). 하지만 데이터의 개수 N 이 충분히 커진다면 posterior pdf $p(A|x)$ 는 점차 \bar{x} 중심의 뾰족한 모양으로 변하며 가우시안에 근사하게 된다. 그렇게 되면 MMSE는 사전 분포의 영향을 점차 덜 받게 되어 MMSE는 \bar{x} 에 근접하게 된다.



10.2 Choosing a Prior PDF

이전 섹션에서 보았듯이 한 번 prior pdf가 정해지면 MMSE는 (352)을 통해 바로 구할 수 있다. 고전적인 방법에서는 MVUE를 구하는 closed form solution이 존재하는 반면 베이지안 추정 방법에서는 여전히 $p(A|x)$ 를 closed form solution으로 결정할 수 있는지 여부가 걸림돌로 남아있다.

이전 예제에서는 (354) 식의 posterior pdf $p(A|x)$ 와 (356) 식의 posterior 평균의 값을 closed form으로 찾을 수 없었고 따라서 이런 상황에서는 MMSE 추정값을 찾기 위해서는 수치적분을 사용해야 한다. **실제 추정 문제에서**는 $p(A|x)$ 를 closed form으로 찾는 것이 중요하기 때문에 이를 가능하게 해주는 prior pdf 선정 방법에 대해 설명한다.

10.2.1 Example 10.1 - DC Level in WGN - Gaussian Prior PDF

이전 예제에서 prior pdf를 아래와 같은 균일 분포(uniform distribution)으로 설정하였다.

$$p(A) = \begin{cases} \frac{1}{2A_0} & |A| \leq A_0 \\ 0 & |A| > A_0 \end{cases} \quad (357)$$

이번에는 $p(A)$ 를 가우시안 분포를 가진다고 가정해보자.

$$p(A) = \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left[-\frac{1}{2\sigma_A^2}(x - \mu_A)^2\right] \quad (358)$$

A 에 대한 두 분포는 매우 다른 표현법을 가지는 것을 알 수 있다. 다음으로 likelihood는 다음과 같다.

$$\begin{aligned} p(\mathbf{x}|\mathbf{A}) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n]\right] \exp\left[-\frac{1}{2\sigma^2}(NA^2 - 2NA\bar{x})\right] \end{aligned} \quad (359)$$

따라서 posterior pdf $p(A|x)$ 는 다음과 같이 전개할 수 있다.

$$\begin{aligned} p(A|x) &= \frac{p(\mathbf{x}|A)p(A)}{\int p(\mathbf{x}|A)p(A)dA} \\ &= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\sqrt{2\pi\sigma_A^2}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]\right]\exp\left[-\frac{1}{2\sigma^2}(NA^2 - 2NA\bar{x})\right]}{\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\sqrt{2\pi\sigma_A^2}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]\right]\exp\left[-\frac{1}{2\sigma^2}(NA^2 - 2NA\bar{x})\right]} \\ &\quad \cdot \frac{\exp\left[-\frac{1}{2\sigma_A^2}(x - \mu_A)^2\right]}{\exp\left[-\frac{1}{2\sigma_A^2}(x - \mu_A)^2\right]}dA \\ &= \frac{\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(NA^2 - 2NA\bar{x}) + \frac{1}{\sigma_A^2}(A - \mu_A)^2\right)\right]}{\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(NA^2 - 2NA\bar{x}) + \frac{1}{\sigma_A^2}(A - \mu_A)^2\right)\right]dA} \\ &= \frac{\exp\left[-\frac{1}{2}Q(A)\right]}{\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}Q(A)\right]dA} \end{aligned} \quad (360)$$

위 식의 분모는 정규화(normalization) 역할을 수행하기 때문에 A 에 종속적이지 않은 것을 알 수 있다. 그리고 exponential 항 내부에 있는 $Q(A)$ 항은 A 에 대한 2차식 형태(quadratic form)인 것을 알 수 있다. **따라서 prior pdf, likelihood가 모두 가우시안이면 posterior pdf 또한 x에 대한 평균과 분산을 가지는 가우시안 분포임을 알 수 있다.** $Q(A)$ 를 자세히 전개해보면 다음과 같다.

$$\begin{aligned} Q(A) &= \frac{N}{\sigma^2}A^2 - \frac{2NA\bar{x}}{\sigma^2} + \frac{A^2}{\sigma_A^2} - \frac{2\mu_AA}{\sigma_A^2} + \frac{\mu_A^2}{\sigma_A^2} \\ &= \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)A^2 - 2\left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu_A}{\sigma_A^2}\right)A + \frac{\mu_A^2}{\sigma_A^2} \end{aligned} \quad (361)$$

Posterior의 평균과 분산을 다음과 같이 정의하자.

$$\begin{aligned}\sigma_{A|x}^2 &= \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \\ \mu_{A|x} &= \left(\frac{N}{\sigma^2} \bar{x} + \frac{\mu_A}{\sigma_A^2} \right) \sigma_{A|x}^2\end{aligned}\tag{362}$$

(361) 식에 위를 대입하면 다음과 같은 2차식 형태로 변한다.

$$\begin{aligned}Q(A) &= \frac{1}{\sigma_{A|x}^2} (A^2 - 2\mu_{A|x}A + \mu_{A|x}^2) - \frac{\mu_{A|x}^2}{\sigma_{A|x}^2} + \frac{\mu_A^2}{\sigma_A^2} \\ &= \frac{1}{\sigma_{A|x}^2} (A^2 - \mu_{A|x}^2)^2 - \frac{\mu_{A|x}^2}{\sigma_{A|x}^2} + \frac{\mu_A^2}{\sigma_A^2}\end{aligned}\tag{363}$$

따라서 $p(A|x)$ 는 다음과 같이 쓸 수 있다.

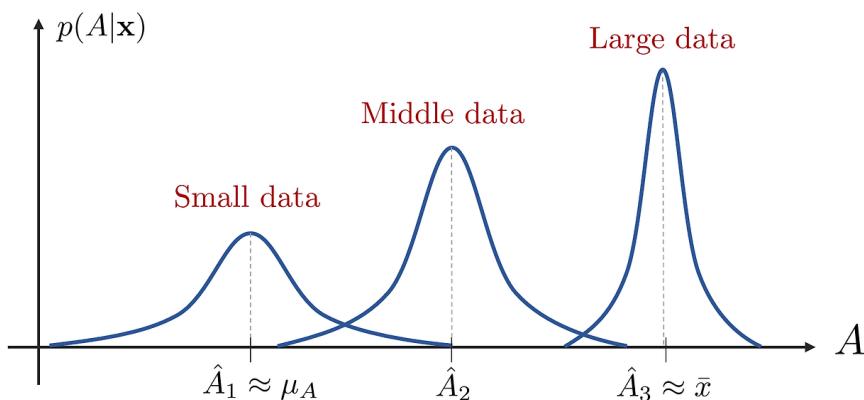
$$\begin{aligned}p(A|x) &= \frac{\exp \left[-\frac{1}{2\sigma_{A|x}^2} (A^2 - \mu_{A|x}^2)^2 \right] \exp \left[-\frac{1}{2} \left(\frac{\mu_A^2}{\sigma_A^2} - \frac{\mu_{A|x}^2}{\sigma_{A|x}^2} \right) \right]}{\int_{-\infty}^{\infty} \exp \left[-\frac{1}{2\sigma_{A|x}^2} (A^2 - \mu_{A|x}^2)^2 \right] \exp \left[-\frac{1}{2} \left(\frac{\mu_A^2}{\sigma_A^2} - \frac{\mu_{A|x}^2}{\sigma_{A|x}^2} \right) \right] dA} \\ &= \frac{1}{\sqrt{2\pi\sigma_{A|x}^2}} \exp \left[-\frac{1}{2\sigma_{A|x}^2} (A - \mu_{A|x})^2 \right]\end{aligned}\tag{364}$$

앞서 언급한 것처럼 posterior pdf 또한 가우시안 분포를 가짐을 알 수 있다. MMSE 추정값을 구해보면 다음과 같다.

$$\begin{aligned}\hat{A} &= \mathbb{E}(A|x) \\ &= \mu_{A|x} \\ &= \frac{\frac{N}{\sigma^2} \bar{x} + \frac{\mu_A}{\sigma_A^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \\ &= \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \bar{x} + \frac{\frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}} \mu_A \\ &= \alpha \bar{x} + (1 - \alpha) \mu_A\end{aligned}\tag{365}$$

$$- \alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}$$

α 는 $0 \leq \alpha \leq 1$ 의 값을 가지는 가중치 값이다. Prior pdf를 가우시안 분포로 가정하면 위와 같이 MMSE 추정값을 closed form으로 계산할 수 있다. MMSE 추정값 \hat{A} 를 해석해보면 데이터가 거의 없는 경우 ($\sigma_A^2 \ll \sigma^2/N$), α 값은 작은 값을 가지며 $\hat{A} \approx \mu_A$ 가 되는 것을 알 수 있다. 하지만 데이터가 충분히 관측된 경우 ($\sigma_A^2 \gg \sigma^2/N$), α 값은 1에 근접하여 $\hat{A} \approx \bar{x}$ 가 됨을 알 수 있다. 이렇듯 가중치 값은 사전 지식의 신뢰도와 관측된 데이터 신뢰도 사이의 균형을 잡아주는 값이다. 아래 그림에서 보다시피 N 값이 증가할수록 posterior pdf는 점점 뾰족해진다.



이러한 현상은 posterior pdf의 분산과 관련이 있다.

$$\begin{aligned}\text{var}(A|\mathbf{x}) &= \sigma_{A|x}^2 \\ &= \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A}}\end{aligned}\tag{366}$$

위 식에서 보다시피 N 이 커질수록 $\text{var}(A|\mathbf{x})$ 는 작아지는 것을 알 수 있다. 만약 prior pdf가 주어지지 않았다면 $\sigma_A^2 \rightarrow \infty$ 로 세팅한 것과 동일한 효과를 보이며 $\hat{A} \rightarrow \bar{x}$ 가 된다. 따라서 고전적인 추정 문제와 동일한 추정값을 얻게 된다.

마지막으로 앞서 주장했던 'prior pdf를 사용하면 더 좋은 추정값을 얻을 수 있다'는 사실을 증명해보자. BMSE를 다시 쓰면 다음과 같다.

$$\text{Bmse}(\hat{A}) = \mathbb{E}[(A - \hat{A})^2]\tag{367}$$

기대값의 정의에 따라 전개해보자.

$$\begin{aligned}\text{Bmse}(\hat{A}) &= \int \int (A - \hat{A})^2 p(\mathbf{x}, A) d\mathbf{x} dA \\ &= \int \int (A - \hat{A})^2 p(A|\mathbf{x}) dA p(\mathbf{x}) d\mathbf{x}\end{aligned}\tag{368}$$

MMSE 추정값은 $\hat{A} = \mathbb{E}(A|\mathbf{x})$ 인 것을 알고 있으므로 이를 대입한다.

$$\begin{aligned}\text{Bmse}(\hat{A}) &= \int \int (A - \mathbb{E}(A|\mathbf{x}))^2 p(A|\mathbf{x}) dA p(\mathbf{x}) d\mathbf{x} \\ &= \int \text{var}(A|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}\end{aligned}\tag{369}$$

위 식에서 보다시피 BMSE는 posterior pdf의 분산이 \mathbf{x} 의 확률 분포와 곱해진 값으로 해석할 수 있다. $\text{var}(A|\mathbf{x})$ 는 \mathbf{x} 와 독립이므로 아래와 같이 쓸 수 있다.

$$\begin{aligned}\text{Bmse}(\hat{A}) &= \int \sigma_{A|x}^2 p(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A}} \\ &= \frac{\sigma^2}{N} \left(\frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \right) \\ &< \frac{\sigma^2}{N}\end{aligned}\tag{370}$$

따라서 BMSE 값은 사전 지식(prior knowledge)가 없는 MMSE 값($\frac{\sigma^2}{N}$)보다 항상 작음을 알 수 있다 ($\sigma_A^2 \rightarrow \infty$ 경우 제외). 결론적으로 베이지안 추정 방법을 통해 prior pdf를 사용하면 항상 추정값의 성능을 향상시킬 수 있다.

10.3 Properties of the Gaussian PDF

이번 섹션에서는 다변량 가우시안 pdf의 성질을 자세히 알아보자. 우선 이변량 가우시안 분포 함수(bivariate Gaussian pdf)에 대해 먼저 알아 본 후 다변량 가우시안 분포 함수(multivariate Gaussian pdf)에 대해 순차적으로 알아본다.

두 가우시안 확률 변수(random variable) $[x, y]^T$ 이 주어졌다고 하자. 두 확률 변수에 대한 결합 분포(joint pdf)는 다음과 같다.

$$p(x, y) = \frac{1}{2\pi \det^{\frac{1}{2}}(\mathbf{C})} \exp \left[-\frac{1}{2} \begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix} \right]\tag{371}$$

위 식은 이변량 가우시안 분포라고도 불린다. 평균과 공분산은 다음과 같은 벡터의 형태를 띤다.

$$\begin{aligned}\mathbb{E} \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) &= \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(y) \end{bmatrix} \\ \mathbf{C} &= \begin{bmatrix} \text{var}(x) & \text{cov}(x, y) \\ \text{cov}(y, x) & \text{var}(y) \end{bmatrix}\end{aligned}\tag{372}$$

$p(x), p(y)$ 의 주변 분포(marginal pdf) 또한 가우시안 분포가 된다. 이는 아래 식을 통해 알 수 있다.

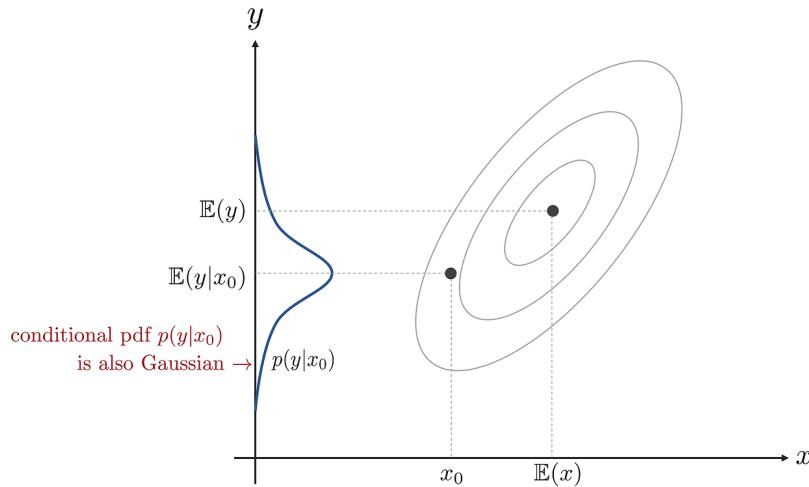
$$\begin{aligned} p(x) &= \int_{-\infty}^{\infty} p(x, y) dy = \frac{1}{\sqrt{2\pi\text{var}(x)}} \exp \left[-\frac{1}{2\text{var}(x)}(x - \mathbb{E}(x))^2 \right] \\ p(y) &= \int_{-\infty}^{\infty} p(x, y) dx = \frac{1}{\sqrt{2\pi\text{var}(y)}} \exp \left[-\frac{1}{2\text{var}(y)}(y - \mathbb{E}(y))^2 \right] \end{aligned} \quad (373)$$

$p(x, y)$ 의 공분산에 대한 타원형 등고선(contour)은 아래 값들이 변하지 않는 이상 일정한 값을 지닌다.

$$\begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix}^\top \mathbf{C}^{-1} \begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix} \quad (374)$$

만약 x_0 데이터가 관측되었다고 하면 y 에 대한 조건부 확률은 다음과 같다.

$$\begin{aligned} p(y|x_0) &= \frac{p(x_0, y)}{p(x_0)} \\ &= \frac{p(x_0, y)}{\int_{-\infty}^{\infty} p(x_0, y) dy} \end{aligned} \quad (375)$$



y 에 대한 조건부 확률 $p(y|x_0)$ 는 분모항에 의해 정규화되어 크기가 1을 가진 가우시안 분포가 된다. 그리고 $p(x_0, y)$ 는 (371)식에서 본 것처럼 y 에 2차식 형태가 되므로 y 에 대한 가우시안 분포를 가짐을 알 수 있다. 결론적으로 $p(x, y)$ 가 서로 결합 가우시안 분포를 이루는 경우(jointly gaussian) $p(y)$ 와 $p(y|x)$ 는 모두 가우시안 분포를 따르는 것을 알 수 있다. 보다 자세한 유도 과정은 Appendix 10A를 참조하면 된다.

10.3.1 Theorem 10.1 (Conditional PDF of Bivariate Gaussian)

두 확률 변수 x, y 가 다음과 같은 평균과 공분산을 가지는 결합 분포를 가진다고 하자.

$$p(x, y) = \frac{1}{2\pi \det^{\frac{1}{2}}(\mathbf{C})} \exp \left[-\frac{1}{2} \begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix}^\top \mathbf{C}^{-1} \begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix} \right] \quad (376)$$

$$\begin{aligned} \text{where, } \mathbb{E} \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) &= \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(y) \end{bmatrix} \\ \mathbf{C} &= \begin{bmatrix} \text{var}(x) & \text{cov}(x, y) \\ \text{cov}(y, x) & \text{var}(y) \end{bmatrix} \end{aligned} \quad (377)$$

이 때, 조건부 pdf $p(y|x)$ 또한 가우시안 분포를 따른다. 자세한 유도 과정은 Appendix 10A를 참고하면 된다.

$$\begin{aligned} \mathbb{E}(y|x) &= \mathbb{E}(y) + \frac{\text{cov}(x, y)}{\text{var}(x)}(x - \mathbb{E}(x)) \\ \text{var}(y|x) &= \text{var}(y) - \frac{\text{cov}^2(x, y)}{\text{var}(x)} \end{aligned} \quad (378)$$

조건부 pdf는 다음과 같이 해석할 수 있다. x 가 관찰되기 전까지 확률 변수 y 는 $p(y) \sim \mathcal{N}(\mathbb{E}(y), \text{var}(y))$ 를 따르고 있다. x 가 관찰되면 y 는 평균과 분산이 (378)처럼 약간 변한 가우시안 분포가 된다. 이 때, 두 확률 변수는 서로 종속적이라고 가정한다($\text{cov}(x, y) \neq 0$). y 에 대한 조건부 pdf는 점점 더 불확실성이 줄어들면서 뾰족한 모양이 된다.

y 의 불확실성이 줄어든다는 것을 수학적으로 확인하기 위해 (378) 분산식을 다시 한 번 살펴보자.

$$\begin{aligned}\text{var}(y|x) &= \text{var}(y) \left[1 - \frac{\text{cov}^2(x, y)}{\text{var}(x)\text{var}(y)} \right] \\ &= \text{var}(y)(1 - \rho^2)\end{aligned}\quad (379)$$

$$- \rho = \frac{\text{cov}(x, y)}{\text{var}(x)\text{var}(y)}$$

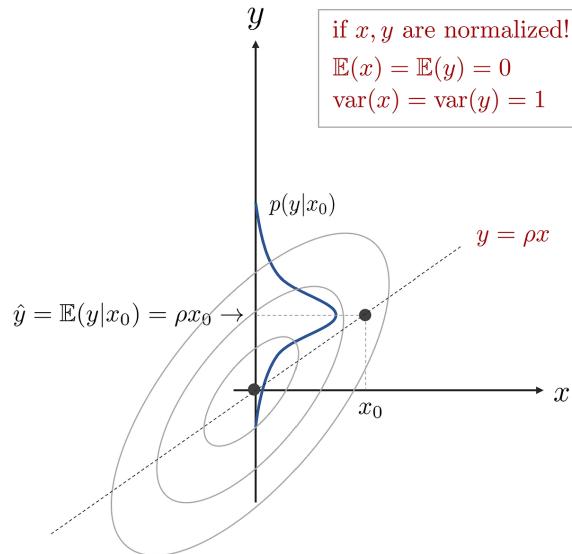
ρ 는 두 변수의 상관계수(cross-correlation coefficient)라고 하며 $|\rho| \leq 1$ 의 조건을 만족한다. 이전 섹션에서 $\mathbb{E}(y|x)$ 는 y 에 대한 MMSE 추정값임을 배웠다. 따라서 (378)의 평균은 다음과 같이 쓸 수 있다.

$$\hat{y} = \mathbb{E}(y) + \frac{\text{cov}(x, y)}{\text{var}(x)}(x - \mathbb{E}(x)) \quad (380)$$

$\mathbb{E}(y)$ 를 이항하고 양변에 $\sqrt{\text{var}(y)}$ 를 나누어주고 $\text{var}(x) \rightarrow \sqrt{\text{var}(x)}\sqrt{\text{var}(x)}$ 로 나누어 주면 다음과 같은 정규화된 형태를 얻을 수 있다. 이 때 평균은 0이고 분산은 1을 가진다.

$$\begin{aligned}\frac{\hat{y} - \mathbb{E}(y)}{\sqrt{\text{var}(y)}} &= \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} \frac{x - \mathbb{E}(x)}{\sqrt{\text{var}(x)}} \\ \therefore \hat{y}_n &= \rho x_n\end{aligned}\quad (381)$$

위 식에서 보다시피 ρ 는 정규화된 관측값 x_n 의 크기를 조절하여 정규화된 MMSE 추정값 \hat{y}_n 의 값을 구하는데 사용된다. 만약 두 확률 변수가 이미 정규화된 경우 ($\mathbb{E}(x) = \mathbb{E}(y) = 0$, $\text{var}(x) = \text{var}(y) = 1$)면 분산 타원형 등고선(contour)은 아래 그림과 같이 얻어진다.



조건부 pdf의 피크(peak) 값은 $y = \rho x$ 직선 위에서 얻을 수 있다. 따라서 MMSE 추정값 $\mathbb{E}(y|x)$ 은 두 확률 변수의 상관 관계를 이용하여 추정값을 평가한다는 것을 알 수 있다. BMSE를 다시 쓰면 다음과 같다.

$$\begin{aligned}\text{Bmse}(\hat{A}) &= \int \text{var}(y|x)p(\mathbf{x})d\mathbf{x} \\ &= \text{var}(y|x) \\ &= \text{var}(y)(1 - \rho^2)\end{aligned}\quad (382)$$

위 식에서 보다시피 posterior의 분산은 x 와 독립적이기 때문에 추정값의 성능은 상관 계수 ρ 에만 의존적인 것을 알 수 있다. 상관 관계는 이름처럼 결국 두 확률 변수 x, y 의 통계적 상관성과 관련이 있으므로 추정값의 성능은 결국 두 확률 변수의 상관성과 관련이 있음을 알 수 있다.

지금까지 이변량 가우시안(bivariate Gaussian)에 대해 알아봤다면 이제는 이를 일반화시켜서 두 벡터 확률 변수 $[\mathbf{x}^\top, \mathbf{y}^\top]^\top$ 이 주어졌을 때 둘 사이의 다변량 가우시안 분포 함수(multivariate Gaussian pdf)에 대해 알아보자.

10.3.2 Theorem 10.2 (Conditional PDF of Multivariate Gaussian)

두 벡터 확률 변수 $\mathbf{x} \in \mathbb{R}^{k \times 1}$ 와 $\mathbf{y} \in \mathbb{R}^{l \times 1}$ 가 주어졌을 때 둘의 joint pdf는 다음과 같이 나타낼 수 있다.

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^{\frac{k+l}{2}} \det^{\frac{1}{2}}(\mathbf{C})} \exp \left[-\frac{1}{2} \begin{bmatrix} \mathbf{x} - \mathbb{E}(\mathbf{x}) \\ \mathbf{y} - \mathbb{E}(\mathbf{y}) \end{bmatrix}^\top \mathbf{C}^{-1} \begin{bmatrix} \mathbf{x} - \mathbb{E}(\mathbf{x}) \\ \mathbf{y} - \mathbb{E}(\mathbf{y}) \end{bmatrix} \right] \quad (383)$$

$$\text{where, } \mathbb{E} \left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right) = \begin{bmatrix} \mathbb{E}(\mathbf{x}) \\ \mathbb{E}(\mathbf{y}) \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = \begin{bmatrix} k \times k & k \times l \\ l \times k & l \times l \end{bmatrix} \quad (384)$$

이 때, 조건부 pdf $p(\mathbf{y}|\mathbf{x})$ 또한 가우시안 분포를 따른다. 자세한 유도 과정은 Appendix 10A를 참고하면 된다.

$$\boxed{\begin{aligned} \mathbb{E}(\mathbf{y}|\mathbf{x}) &= \mathbb{E}(\mathbf{y}) + \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} (\mathbf{x} - \mathbb{E}(\mathbf{x})) \\ \mathbf{C}_{y|x} &= \mathbf{C}_{yy} - \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \end{aligned}} \quad (385)$$

위 식에서 posterior의 공분산은 x와 독립적임에 주목하자. 이러한 특성은 추후 섹션에서 유용하게 사용된다. 앞서 설명하였듯이 벡터 케이스에서도 \mathbf{x}, \mathbf{y} 가 서로 결합 분포를 이루면(joint gaussian), $p(\mathbf{y})$ 와 $p(\mathbf{y}|\mathbf{x})$ 는 둘 다 가우시안 분포를 따른다.

10.4 Bayesian Linear Model

다음과 같은 데이터 모델이 주어졌다고 하자.

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1 \quad (386)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

그리고 베이지안 철학에 따라 $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$ 라고 하자. 이를 벡터 형태로 표현하면 다음과 같다.

$$\mathbf{x} = \mathbf{A}\mathbf{w} + \mathbf{w} \quad (387)$$

이는 파라미터 A 가 확률 변수라는 것만 제외하면 앞선 챕터에서 말한 선형 모델과 동일하다. 위 식을 다음과 같이 일반적인 형태로 표현하면 다음과 같다.

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \quad (388)$$

- $\mathbf{x} \in \mathbb{R}^{N \times 1}$
- $\mathbf{H} \in \mathbb{R}^{N \times p}$
- $\boldsymbol{\theta} \in \mathbb{R}^{p \times 1} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$
- $\mathbf{w} \in \mathbb{R}^{N \times 1} \sim \mathcal{N}(0, \mathbf{C}_w)$: $\boldsymbol{\theta}$ 와 독립적

위와 같은 선형 모델을 베이지안 일반 선형 모델(Bayesian general linear model)이라고 한다. 위 식에서 $p(\boldsymbol{\theta}|\mathbf{x})$ 의 값이 어떻게 나올지 궁금해할 수 있다. 이는 베이지안 추정 방법에 따라 두 확률 변수 $\mathbf{x}, \boldsymbol{\theta}$ 가 결합 분포를 이루므로(jointly gaussian) $p(\boldsymbol{\theta}|\mathbf{x})$ 또한 가우시안 분포를 가짐을 알 수 있다. $\mathbf{z} = [\mathbf{x}^\top, \boldsymbol{\theta}^\top]^\top$ 과 같이 정의했을 때 이를 자세히 보면 다음과 같다.

$$\begin{aligned} \mathbf{z} &= \begin{bmatrix} \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{H} & \mathbf{I}_N \\ \mathbf{I}_p & \mathbf{0}_N \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{w} \end{bmatrix} \end{aligned} \quad (389)$$

- $\mathbf{I}_N \in \mathbb{R}^{N \times N}$
- $\mathbf{I}_p \in \mathbb{R}^{p \times p}$
- $\mathbf{0}_N \in \mathbb{R}^{N \times N}$

$\boldsymbol{\theta}$ 와 \mathbf{w} 은 서로 독립이며 둘 중 하나는 가우시안 분포를 따르기 때문에 둘은 결합 분포를 이룸(jointly gaussian)을 알 수 있다. 추가적으로 \mathbf{z} 는 두 확률 변수의 선형 변환인기 때문에 이 역시 가우시안 분포를 따름을 알 수 있다. 따라서 Theorem 10.2를 바로 적용할 수 있다. $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ 이고 $\mathbf{y} = \boldsymbol{\theta}$ 이 된다.

$$\begin{aligned} \mathbb{E}(\mathbf{x}) &= \mathbb{E}(\mathbf{H}\boldsymbol{\theta} + \mathbf{w}) = \mathbf{H}\mathbb{E}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\mu}_\theta \\ \mathbb{E}(\mathbf{y}) &= \mathbb{E}(\boldsymbol{\theta}) = \boldsymbol{\mu}_\theta \end{aligned} \quad (390)$$

공분산 \mathbf{C}_{xx} 은 다음과 같다.

$$\begin{aligned}
\mathbf{C}_{xx} &= \mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^\top] \\
&= \mathbb{E}[(\mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{H}\boldsymbol{\mu}_\theta)(\mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{H}\boldsymbol{\mu}_\theta)^\top] \\
&= \mathbb{E}[(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \mathbf{w})(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \mathbf{w})^\top] \\
&= \mathbf{H}\mathbb{E}[(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)^\top]\mathbf{H}^\top + \mathbb{E}[\mathbf{w}\mathbf{w}^\top] \\
&= \mathbf{H}\mathbf{C}_\theta\mathbf{H}^\top + \mathbf{C}_w
\end{aligned} \tag{391}$$

$\boldsymbol{\theta}$ 와 \mathbf{w} 가 서로 독립적이라는 사실에 유의하여 \mathbf{C}_{yx} 를 구해보면 다음과 같다.

$$\begin{aligned}
\mathbf{C}_{yx} &= \mathbb{E}[(\mathbf{y} - \mathbb{E}(\mathbf{y}))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^\top] \\
&= \mathbb{E}[(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \textcolor{red}{w})^\top] \\
&= \mathbb{E}[(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta))^\top] \\
&= \mathbf{C}_\theta\mathbf{H}^\top
\end{aligned} \tag{392}$$

10.4.1 Theorem 10.3 (Posterior PDF for the Bayesian General Linear Model)

관측 데이터 \mathbf{x} 가 다음과 같은 선형 모델로 표현될 수 있는 경우

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{393}$$

- $\mathbf{x} \in \mathbb{R}^{N \times 1}$
- $\mathbf{H} \in \mathbb{R}^{N \times p}$
- $\boldsymbol{\theta} \in \mathbb{R}^{p \times 1} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$
- $\mathbf{w} \in \mathbb{R}^{N \times 1} \sim \mathcal{N}(0, \mathbf{C}_w) : \boldsymbol{\theta}$ 와 독립적

Posterior pdf $p(\boldsymbol{\theta}|\mathbf{x})$ 의 평균과 공분산은 다음과 같이 나타낼 수 있다.

$$\boxed{
\begin{aligned}
\mathbb{E}(\boldsymbol{\theta}|\mathbf{x}) &= \boldsymbol{\mu}_\theta + \mathbf{C}_\theta\mathbf{H}^\top(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^\top + \mathbf{C}_w)^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta) \\
\mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} &= \mathbf{C}_\theta - \mathbf{C}_\theta\mathbf{H}^\top(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^\top + \mathbf{C}_w)^{-1}\mathbf{H}\mathbf{C}_\theta
\end{aligned} \tag{394}
}$$

고전적인 추정 문제의 선형 모델과 비교해보면 $(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^\top + \mathbf{C}_w)$ 의 역행렬을 구하기 위해 \mathbf{H} 행렬은 반드시 full rank가 아니어도 된다. 추후 섹션에서 설명의 편의를 위해 (394)의 변형된 버전에 대하여 설명한다.

$$\boxed{
\begin{aligned}
\mathbb{E}(\boldsymbol{\theta}|\mathbf{x}) &= \boldsymbol{\mu}_\theta + (\mathbf{C}_\theta^{-1} + \mathbf{H}^\top\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{C}_w^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta) \\
\mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} &= (\mathbf{C}_\theta^{-1} + \mathbf{H}^\top\mathbf{C}_w^{-1}\mathbf{H})^{-1} \\
\mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}}^{-1} &= \mathbf{C}_\theta^{-1} + \mathbf{H}^\top\mathbf{C}_w^{-1}\mathbf{H}
\end{aligned} \tag{395}
}$$

위 식에서 보다시피 만약 사전 지식(prior knowledge)가 주어지지 않는다면 ($\boldsymbol{\mu}_\theta \rightarrow 0, \mathbf{C}_\theta^{-1} \rightarrow 0$) , MMSE 추정값은 $\hat{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}|\mathbf{x}) = (\mathbf{H}^\top\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{C}_w^{-1}\mathbf{x}$ 가 되어서 선형 모델의 MVUE와 동일한 공식을 얻는다.

10.5 Bayesian Estimation for Deterministic Parameters

베이지안 추정 방법은 엄밀히 말해서 파라미터 $\boldsymbol{\theta}$ 랜덤한 확률 변수일 때만 적용되지만, 실제로는 결정론적(deterministic) 파라미터 추정에도 종종 사용된다. 이는 베이지안 가정을 사용하여 추정값을 도출하고 마치 $\boldsymbol{\theta}$ 가 랜덤 변수가 아닌 것처럼 그 추정값을 사용하는 것을 의미한다. 이러한 상황은 MVUE가 존재하지 않을 때 발생할 수 있다. 예를 들어, 우리가 분산 측면에서 다른 추정기들보다 일관되게 우수한 불편추정값(unbiased estimator)을 찾지 못할 수도 있지만 베이지안 프레임워크 내에서 MMSE는 항상 존재하고 이는 다양한 $\boldsymbol{\theta}$ 값들에 대해 평균적으로 잘 작동하는 추정값을 제공한다. 하지만 평균적으로 잘 작동한다는 것일뿐 가끔씩 안 좋은 성능을 낼 수도 있음을 유의하자.

Example 10.1 문제를 다시 보자. (365) 식을 보면 다음과 같다.

$$\begin{aligned}
\hat{A} &= \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\bar{x} + \frac{\frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}}\mu_A \\
&= \alpha\bar{x} + (1 - \alpha)\mu_A
\end{aligned} \tag{396}$$

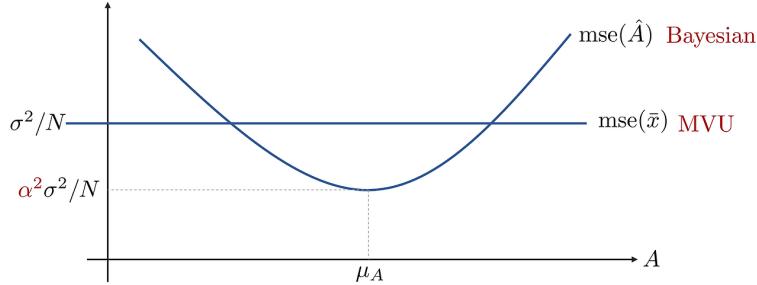
$$- \alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}$$

α 는 $0 \leq \alpha \leq 1$ 의 값을 가지는 가중치 값이다. 만약 A 가 결정론적(deterministic) 파라미터이라고 가정해보자. MSE에 위 식을 대입하면 다음과 같다.

$$\begin{aligned}
\text{mse}(\hat{A}) &= \text{var}(\hat{A}) + b^2(\hat{A}) \\
&= \alpha^2 \text{var}(\bar{x}) + [\alpha A + (1 - \alpha)\mu_A - A]^2 \\
&= \alpha^2 \frac{\sigma^2}{N} + (1 - \alpha)^2 (A - \mu_A)^2
\end{aligned} \tag{397}$$

- $b(\hat{A}) = \mathbb{E}(\hat{A}) - A$: 편향(bias)

위 식에서 보다시피 베이지안 추정값을 사용하면 $0 \leq \alpha^2 \leq 1$ 에 의해 분산을 줄일 수 있지만(빨강), 편향(bias) 성분이 상당히 증가하는 것을 볼 수 있다(파랑).



Bayesian is better only if A is closer to the prior mean μ_A !

위 그림에서 보다시피 베이지안 추정값은 파라미터 A 가 prior mean μ_A 근처에 있을 때만 MVUE \bar{x} 보다 낮은 MSE 값을 가진다. 그렇지 않으면 오히려 성능이 떨어진다. 이렇듯 베이지안 추정 문제에서는 다른 모든 추정값들보다 MSE가 낮은 하나의 추정값은 존재하지 않지만 '평균적으로' 좋은 추정값은 존재한다. BMSE를 구해보면 다음과 같다.

$$\begin{aligned}
\text{Bmse}(\hat{A}) &= \mathbb{E}_A[\text{mse}(\hat{A})] \\
&= \alpha^2 \frac{\sigma^2}{N} + (1 - \alpha)^2 \mathbb{E}_A[(A - \mu_A)^2] \\
&= \alpha^2 \frac{\sigma^2}{N} + (1 - \alpha)^2 \sigma_A^2 \\
&= \frac{\sigma^2}{N} \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \\
&< \frac{\sigma^2}{N} = \text{Bmse}(\bar{x})
\end{aligned} \tag{398}$$

따라서 MSE는 조건적으로 베이지안 추정값 \hat{A} 이 MVUE \bar{x} 보다 높을 수 있지만 BMSE는 \hat{A} 이 항상 작은 것을 알 수 있다. 결국 베이지안 추정 방법을 통해 파라미터 A 의 사전 지식을 안다는 것은 평균적으로 좋은 성능을 발휘하지만 가끔씩 안 좋은 prior를 추정하게 되면 MVUE보다 성능이 낮을 수 있음을 의미한다.

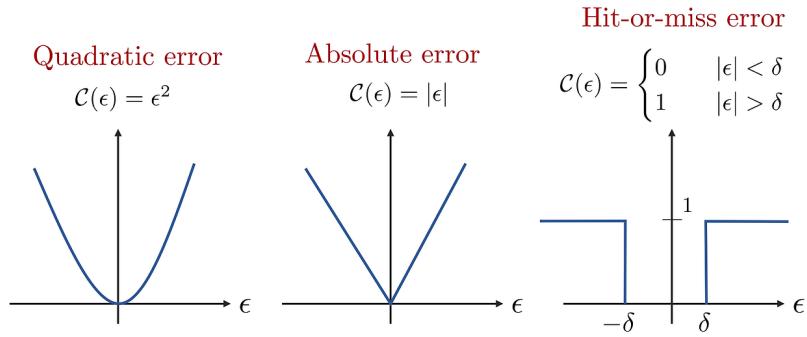
11 General Bayesian Estimators

이전 챕터에서는 베이지안 추정 방법을 사용하여 파라미터를 추정하는 방법에 대해 배웠다. 이번 챕터에서는 일반적인 베이지안 추정값들과 그들의 성질에 대해 배운다. 이를 위해 Bayes risk에 대해 설명한다. 그리고 Bayes risk를 최소화함으로써 다양한 추정값들을 얻는 내용에 대해 배운다. 주목할만한 추정값은 MMSE 추정값과 maximum a posteriori 추정값이다. 마지막으로 베이지안 추정값의 성능에 대해 평가한다.

11.1 Risk Functions

이전 챕터에서 우리는 $\mathbb{E}[(\theta - \hat{\theta})^2]$ 의 값을 최소화하는 MMSE 추정값에 대해 배웠다. 예를 $\epsilon = \theta - \hat{\theta}$ 라고 정의하자. ϵ 은 x 와 θ 에 대한 추정값의 에러를 의미한다. 그리고 비용 함수(cost function)를 $\mathcal{C}(\epsilon)$ 로 정의하자. 비용 함수의 기대값은 Bayes risk \mathcal{R} 이라고 정의하자. Bayes risk \mathcal{R} 은 주어진 추정값의 성능을 평가해주는 criterion 함수이다. 지금까지 정의한 내용을 정리하면 다음과 같다.

$$\mathcal{R} = \mathbb{E}[\mathcal{C}(\epsilon)] \tag{399}$$



위 그림과 같이 세 가지 서로 다른 비용 함수를 알아보자. Quadratic 비용함수는 에러의 제곱을 적용한 형태이다.

$$C(\epsilon) = \epsilon^2 \quad \cdots \text{quadratic error} \quad (400)$$

비용 함수가 위와 같은 경우 Bayes risk \mathcal{R} 는 MSE가 된다. 비용 함수는 2차식 이외에도 다양한 형태가 될 수 있다. 다음으로 absolute 비용함수에 대해 알아보면 다음과 같다. 이는 에러에 절대값을 적용한 형태이다.

$$C(\epsilon) = |\epsilon| \quad \cdots \text{absolute error} \quad (401)$$

마지막 비용 함수는 hit-or-miss 비용 함수라고 불리며 δ 보다 작은 에러 값은 반영을 하지 않고 δ 보다 큰 값들은 항상 1로만 반영하는 비용 함수이다.

$$C(\epsilon) = \begin{cases} 0 & |\epsilon| < \delta \\ 1 & |\epsilon| > \delta \end{cases} \quad \cdots \text{hit-or-miss error} \quad (402)$$

δ 는 항상 0보다 크다. 방금 소개한 세 가지 비용 함수는 양의 에러 또는 음의 에러를 구분하지 않고 전부 동일하게 처리하는 공통점이 있다. 물론 이러한 가정은 실제 추정 문제에서는 대부분 참이 아니다.

Bayes risk \mathcal{R} 에 대한 최적의 추정값을 얻기 위해 이를 자세히 전개하면 다음과 같다.

$$\begin{aligned} \mathcal{R} &= \mathbb{E}(C(\epsilon)) \\ &= \int \int C(\theta - \hat{\theta}) p(\mathbf{x}, \theta) d\mathbf{x} d\theta \\ &= \int \left[\int C(\theta - \hat{\theta}) p(\theta | \mathbf{x}) d\theta \right] p(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (403)$$

챕터 10에서 다룬 내용과 동일하게 위 브라켓 안에 있는 값만 최소화하면 된다. 브라켓 안의 \mathbf{x} 값은 주어진 상수가 되고 $\hat{\theta}$ 값은 스칼라 변수가 된다.

Quadratic error case:

$C(\epsilon) = \epsilon^2$ 인 경우 MMSE 추정값을 얻는다는 사실을 이전 챕터에서 이미 설명하였기 때문에 자세한 설명은 생략한다. **MMSE 추정값 $\hat{\theta} = \mathbb{E}(\theta | \mathbf{x})$ 는 posterior pdf의 평균(mean)을 의미한다.**

Absolute error case:

다음으로 $C(\epsilon) = |\epsilon|$ 케이스를 보자. 이를 위 브라켓 안에 대입하면 다음과 같다.

$$\begin{aligned} g(\hat{\theta}) &= \int |\theta - \hat{\theta}| p(\theta | \mathbf{x}) d\theta \\ &= \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta | \mathbf{x}) d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) p(\theta | \mathbf{x}) d\theta \end{aligned} \quad (404)$$

위 식은 라이프니츠 정리(Leibnitz rule)에 의해 다음과 같이 정리된다.

$$\frac{\partial g(\hat{\theta})}{\partial \hat{\theta}} = \int_{-\infty}^{\hat{\theta}} p(\theta | \mathbf{x}) d\theta - \int_{\hat{\theta}}^{\infty} p(\theta | \mathbf{x}) d\theta = 0 \quad (405)$$

$$\therefore \int_{-\infty}^{\hat{\theta}} p(\theta | \mathbf{x}) d\theta = \int_{\hat{\theta}}^{\infty} p(\theta | \mathbf{x}) d\theta \quad \cdots \text{Median of PDF} \quad (406)$$

위 식에서 보다시피 $\hat{\theta}$ 를 기점으로 $-\infty, \infty$ 까지 적분한 확률 크기가 같으려면 $\hat{\theta}$ 는 중앙값에 위치해야만 한다. 따라서 absolute 비용 함수를 사용했을 때 Bayes risk를 최소화하는 값은 posterior pdf의 중앙값(median)이 된다.

Hit-or-miss case:

다음으로 비용 함수가 hit-or-miss인 경우 ($C(\epsilon) = \begin{cases} 0 & |\epsilon| < \delta \\ 1 & |\epsilon| > \delta \end{cases}$)에 대해 알아보자.

$$g(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}-\delta} 1 \cdot p(\theta|\mathbf{x})d\theta + \int_{\hat{\theta}+\delta}^{\infty} 1 \cdot p(\theta|\mathbf{x})d\theta \quad (407)$$

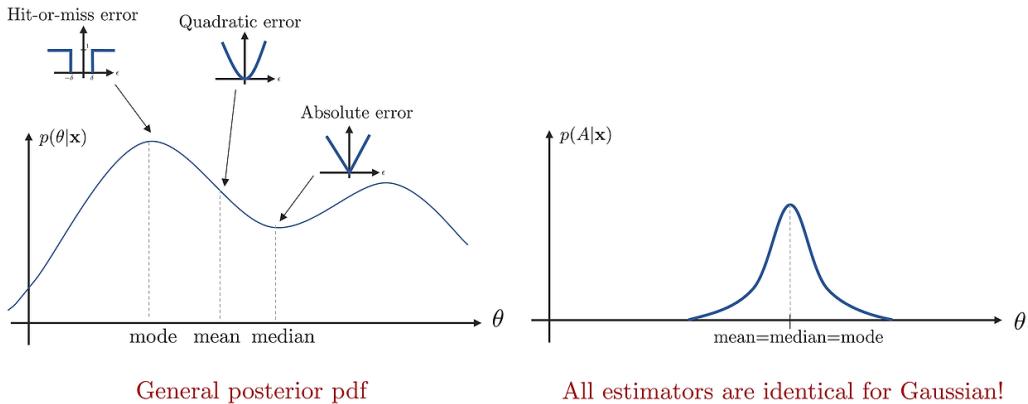
확률의 정의에 의해 $\int_{-\infty}^{\infty} p(\theta|\mathbf{x})d\theta = 1$ 를 만족해야 하므로 위 식은 다음과 같이 정리된다.

$$g(\hat{\theta}) = 1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|\mathbf{x})d\theta \quad (408)$$

위 식은 $\int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|\mathbf{x})d\theta$ 를 최대화시킴으로써 Bayes risk를 최소화시킬 수 있다. 충분히 작은 δ 에 대하여 $\hat{\theta}$ 를 찾는 문제는 가장 큰 $p(\theta|\mathbf{x})$ 값을 찾는 문제와 동치이다. 따라서 hit-or-miss 비용 함수를 사용했을 때 Bayes risk를 최소화하는 값은 posterior pdf의 최빈값(mode)가 된다. 이는 maximum a posteriori(MAP) 추정값이라고도 부른다.

정리하면 Bayes risk를 최소화하는 추정값은 비용 함수에 따라 다르며 quadratic, absolute, hit-or-miss 비용 함수에 대하여 각각 posterior pdf의 mean, median, mode를 의미한다. 만약 pdf가 가우시안 분포를 따른다면 mean, median, mode는 전부 동일한 값을 가진다.

$$p(\theta|\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma_{\theta|x}^2}} \exp\left[-\frac{1}{2\sigma_{\theta|x}^2}(\theta - \mu_{\theta|x})^2\right] \quad (409)$$



11.2 Maximum A Posteriori Estimators

MAP 추정값은 다음과 같이 posterior pdf의 값을 최대화하는 $\hat{\theta}$ 로 정의된다.

$$\hat{\theta} = \arg \max_{\theta} p(\theta|\mathbf{x}) \quad (410)$$

이는 이전 섹션에서 Bayes risk의 비용 함수가 hit-or-miss인 경우 최적의 추정값이 MAP임을 배웠다. Posterior pdf를 베이즈 규칙에 따라 전개하면 다음과 같다.

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} \quad (411)$$

위 식에서 보다시피 MAP은 $p(\mathbf{x}|\theta)p(\theta)$ 를 최대화 하는 것과 동일하다. 이는 MLE 추정값에 prior 정보가 결합된 형태로 해석할 수 있다. 따라서 MAP 추정값은 다음과 같다.

$$\hat{\theta} = \arg \max_{\theta} p(\mathbf{x}|\theta)p(\theta) \quad (412)$$

위 식 우항에 log를 취해도 결과는 변하지 않으므로 다음 식이 성립한다.

$$\hat{\theta} = \arg \max_{\theta} \ln p(\mathbf{x}|\theta) + \ln p(\theta) \quad (413)$$

벡터 파라미터로 MAP 추정값을 확장하기 전에 우선 스칼라 케이스에 대해 알아보자.

11.2.1 Example 11.2 - Exponential PDF

다음과 같은 exponential 확률 분포가 주어졌다고 하자.

$$p(x[n]|\theta) = \begin{cases} \theta \exp(-\theta x[n]) & x[n] > 0 \\ 0 & x[n] \leq 0 \end{cases} \quad (414)$$

- $x[n]$: i.i.d

모든 데이터에 대한 확률 분포는 다음과 같다.

$$p(\mathbf{x}|\theta) = \prod_{n=0}^{N-1} p(x[n]|\theta) \quad (415)$$

Prior pdf도 exponential 확률 분포를 따른다고 가정하자.

$$p(\theta) = \begin{cases} \lambda \exp(-\lambda\theta) & \theta > 0 \\ 0 & \theta \leq 0 \end{cases} \quad (416)$$

(413)에 위 식을 대입해보면 다음과 같다.

$$\begin{aligned} g(\theta) &= \ln p(\mathbf{x}|\theta) + \ln p(\theta) \\ &= \ln \left[\theta^N \exp \left(-\theta \sum_{n=0}^{N-1} x[n] \right) \right] + \ln [\lambda \exp(-\lambda\theta)] \\ &= N \ln \theta - N\theta \bar{x} + \ln \lambda - \lambda\theta \end{aligned} \quad (417)$$

이 때, $\theta > 0$ 이라고 가정한다. 위 식을 미분 후 0으로 놓고 풀면 exponential pdf에 대한 MAP 추정값을 얻을 수 있다.

$$\frac{\partial g(\theta)}{\partial \theta} = \frac{N}{\theta} - N\bar{x} - \lambda = 0 \quad (418)$$

$$\boxed{\hat{\theta} = \frac{1}{\bar{x} + \frac{\lambda}{N}}} \quad (419)$$

만약 데이터가 충분히 많다면 ($N \rightarrow \infty$), $\hat{\theta} \rightarrow \frac{1}{\bar{x}}$ 인 것을 알 수 있다. 또한 사전 지식이 균일 분포에 가까운 모양될 수록 ($\lambda \rightarrow 0$), 역시 $\hat{\theta} \rightarrow \frac{1}{\bar{x}}$ 인 것을 알 수 있다. 이는 베이지안 MLE를 수행한 것과 동일하다. λ 가 0에 가까워질수록 likelihood $p(\mathbf{x}|\theta)$ 값이 prior pdf $p(\theta)$ 보다 커지게 되고 따라서 prior pdf의 영향은 점점 작아진다.

11.2.2 Scalar MAP estimator v.s. vector MAP estimator

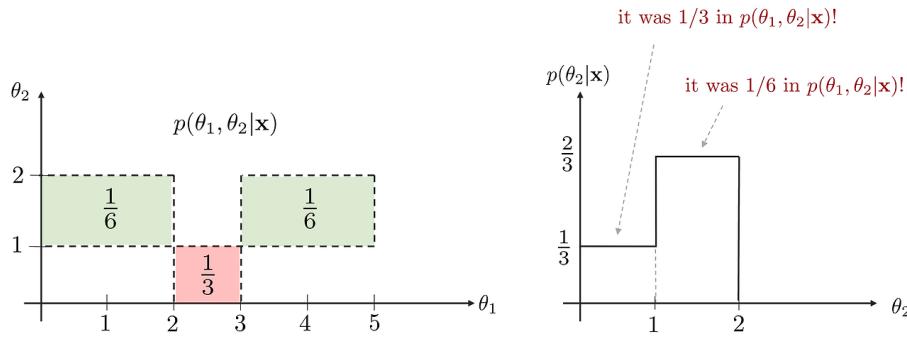
이번 섹션에서는 MAP 추정값을 벡터 파라미터로 확장해보자. 우선 스칼라 MAP 추정값은 다음과 같다.

$$\hat{\theta} = \arg \max_{\theta} p(\theta|\mathbf{x}) \quad (420)$$

스칼라 MAP 추정값의 장점은 오직 하나의 파라미터에 대한 $p(\mathbf{x}|\theta)p(\theta)$ 만 구하면 된다는 것이다. 따라서 여러 파라미터를 결합하는 과정이 생략된다. 벡터 MAP 추정값은 다음과 같이 정의할 수 있다.

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}) \quad (421)$$

벡터 MAP 추정값은 스칼라 버전과는 다른 값이 계산될 수 있다. 스칼라 MAP 추정값은 주어진 데이터 \mathbf{x} 에 대해 하나의 파라미터 θ 의 사후 확률을 최대화하는 값을 찾는다. 반면에, 벡터 MAP 추정값은 여러 파라미터 $\theta_1, \theta_2, \dots, \theta_i$ 의 결합 사후 확률을 최대화하는 값을 찾는다. 결합 분포(joint pdf)를 고려할 때는 두 변수 간의 관계가 중요해지며 이는 벡터 MAP 추정값이 더 복잡한 상황을 반영한 최적값을 계산한다고 볼 수 있다.



예를 들어 $p(\theta_1, \theta_2 | \mathbf{x})$ 가 위 왼쪽 그림과 같은 결합 분포를 가진다고 하자. θ_2 에 대한 벡터 MAP 추정값은 직관적으로 결합 분포에서 확률이 1/3로 제일 높은 $0 < \hat{\theta}_2 < 1$ 구간 안에 존재할 것이다. 하지만 이를 오직 스칼라 $\hat{\theta}_2$ 에 대해서만 추정하게 되면 상황이 달라진다. $p(\theta_1, \theta_2 | \mathbf{x})$ 를 θ_2 에 대하여 주변화(marginalize)해보자.

$$\begin{aligned} p(\theta_2 | \mathbf{x}) &= \int p(\theta_1, \theta_2 | \mathbf{x}) d\theta_1 \\ &= \begin{cases} \int_2^3 \frac{1}{3} d\theta_1 & 0 < \theta_2 < 1 \\ \int_0^2 \frac{1}{6} d\theta_1 + \int_3^5 \frac{1}{6} d\theta_1 & 1 < \theta_1 < 2 \end{cases} \\ &= \begin{cases} \frac{1}{3} & 0 < \theta_2 < 1 \\ \frac{2}{3} & 1 < \theta_2 < 2 \end{cases} \end{aligned} \quad (422)$$

위 그림 오른쪽에서 보는 것과 같이 동일한 구간에 대한 확률이 달라졌다. 따라서 $p(\theta_2 | \mathbf{x})$ 에 대한 MAP 추정값은 확률이 제일 높은 $1 < \hat{\theta}_2 < 2$ 구간 간에 존재하게 된다. 이는 두 파라미터가 결합 분포를 이뤘을 때(jointly gaussian)와 다른 MAP 추정값이다. 따라서 벡터 MAP 추정값은 Bayes risk를 최소화하는 것은 맞지만 더 이상 hit-or-miss 비용 함수를 최소화하는 것은 아니다.

11.2.3 Example 11.5 - Exponential PDF

Example 11.2 예제에서 기존 파라미터 θ 의 비선형 변환된 파라미터 $\alpha = 1/\theta$ 를 추정하는 문제에 대해 생각해보자. MAP 추정값 $\hat{\alpha}$ 은 다음과 같이 가정할 수 있다.

$$\hat{\alpha} = \frac{1}{\hat{\theta}} \quad (423)$$

$\hat{\theta}$ 는 θ 에 대한 MAP 추정값을 의미하며 이전 예제에서 $\hat{\theta} = \frac{1}{\bar{x} + \frac{\lambda}{N}}$ 임을 알고 있다. 그렇다면 $\hat{\alpha}$ 는 다음과 같을까?

$$\hat{\alpha} = \bar{x} + \frac{\lambda}{N} \quad (424)$$

우리는 위 값이 틀렸다는 것을 증명할 것이다. Exponential 확률 분포로 돌아가서 α 를 대입하면 다음과 같다.

$$p(x[n] | \theta) = \begin{cases} \theta \exp(-\theta x[n]) & x[n] > 0 \\ 0 & x[n] \leq 0 \end{cases} \quad (425)$$

$$\Rightarrow p(x[n] | \alpha) = \begin{cases} \frac{1}{\alpha} \exp(-\frac{x[n]}{\alpha}) & x[n] > 0 \\ 0 & x[n] \leq 0 \end{cases} \quad (426)$$

위 식은 θ 에 대한 확률 분포가 아니기 때문에 바로 α 로 변환이 가능하다. 다음으로 prior pdf는 아래와 같다.

$$p(\theta) = \begin{cases} \lambda \exp(-\lambda \theta) & \theta > 0 \\ 0 & \theta \leq 0 \end{cases} \quad (427)$$

하지만 파라미터 θ 는 결정론적(deterministic) 파라미터가 아닌 확률 변수로 보기 때문에 위 식에 $\theta = 1/\alpha$ 와 같이 바로 대입할 수 없다. 확률의 정의를 만족시키기 위해 분모에 α 에 대한 미분 값을 넣어줌으로써 크기를 1로 정규화시킨다.

$$\begin{aligned} p_\alpha(\alpha) &= \frac{p_\theta(\theta(\alpha))}{\left| \frac{d\alpha}{d\theta} \right|} \\ &= \begin{cases} \frac{\lambda \exp(-\lambda/\alpha)}{\alpha^2} & \alpha > 0 \\ 0 & \alpha \leq 0 \end{cases} \end{aligned} \quad (428)$$

지금까지 유도한 식을 바탕으로 MAP 추정값을 찾아보자.

$$\begin{aligned}
 g(\alpha) &= \ln p(\mathbf{x}|\alpha) + \ln p(\alpha) \\
 &= \ln \left[\left(\frac{1}{\alpha} \right)^N \exp \left(-\frac{1}{\alpha} \sum_{n=0}^{N-1} x[n] \right) \right] + \ln \frac{\lambda \exp(-\lambda/\alpha)}{\alpha^2} \\
 &= -N \ln \alpha - N \frac{\bar{x}}{\alpha} + \ln \lambda - \frac{\lambda}{\alpha} - 2 \ln \alpha \\
 &= -(N+2) \ln \alpha - \frac{N\bar{x} + \lambda}{\alpha} + \ln \lambda
 \end{aligned} \tag{429}$$

위 식을 α 에 대하여 미분 후 0으로 놓으면 다음과 같은 MAP 추정값이 구해진다.

$$\frac{dg}{d\alpha} = -\frac{N+2}{\alpha} + \frac{N\bar{x} + \lambda}{\alpha^2} = 0 \tag{430}$$

$$\hat{\alpha} = \frac{N\bar{x} + \lambda}{N+2} \tag{431}$$

이는 앞서 예상한 (424)와 다르다는 것을 알 수 있다. 이를 통해 MAP 추정값은 비선형 변환에 대하여 교환법칙이 성립하지 않는다는 것을 알 수 있다. 이는 MLE의 Invariance 성질과 대조되는 부분이다.

11.3 Performance Description

고전적인 추정 문제에서 우리는 추정값의 평균과 분산에 관심을 두었다. 만약 추정값이 가우시안 분포를 따른다고 하면 우리는 이를 통해 바로 pdf를 구할 수 있었다. 만약 pdf가 참값(true value) 주변에 집중되어 있다면 우리는 그 추정값의 성능이 좋다고 말할 수 있었다.

하지만 베이지안 추정 문제에서 파라미터는 확률 변수(random variable)로 취급되므로 이와 같은 접근법을 사용할 수 없다. 파라미터의 랜덤성으로 인해 각 θ 가 샘플링될 때마다 다른 pdf가 생성된다. 이러한 pdf를 $p(\hat{\theta}|\theta)$ 라고 표기하자. 좋은 추정값의 성능을 위해 θ 와 $\hat{\theta}$ 의 차이는 거의 없어야 하며 이를 평가하기 위해 에러 ϵ 을 다음과 같이 정의한다.

$$\epsilon = \theta - \hat{\theta} \tag{432}$$

좋은 추정값을 얻기 위한 기준은 다음과 같다.

1. 모든 가능한 추정값 $\hat{\theta}$ 은 θ 와 근접할 수록 좋다. 다시 말하면 에러 $\epsilon = \theta - \hat{\theta}$ 가 작아야 한다.
2. 좋은 추정값 $\hat{\theta}$ 을 얻기 위해 $p(\hat{\theta}|\theta)$ 가 θ 주변에 집중적으로 분포해야 한다.
3. 이를 만족하는 추정값은 $\mathbb{E}_{x,\theta}[(\theta - \hat{\theta})^2]$ 를 최소화하는 MMSE 추정값이된다.

베이지안 추정값의 성능을 평가할 때 항상 prior 분포 선정이 문제되므로 에러에 대한 pdf $p(\hat{\theta}|\theta)$ 를 평가하는 것은 적합한 절차이다. 좋은 추정값일 수록 $p(\hat{\theta}|\theta)$ 값이 0 주변에 집중되어 있고 이는 곧 에러가 적다는 의미이다.

위 3번에서 MMSE 추정값은 $\hat{\theta} = \mathbb{E}(\theta|\mathbf{x})$ 이므로 이를 에러에 대입하면 다음과 같다.

$$\epsilon = \theta - \mathbb{E}(\theta|\mathbf{x}) \tag{433}$$

에러의 평균은 다음과 같다.

$$\begin{aligned}
 \mathbb{E}_{x,\theta}(\epsilon) &= \mathbb{E}_{x,\theta}[\theta - \mathbb{E}(\theta|\mathbf{x})] \\
 &= \mathbb{E}_x[\mathbb{E}_{\theta|x}(\theta) - \mathbb{E}_{\theta|x}(\theta|\mathbf{x})] \\
 &= \mathbb{E}_x[\mathbb{E}(\theta|\mathbf{x}) - \mathbb{E}(\theta|\mathbf{x})] \\
 &= 0
 \end{aligned} \tag{434}$$

따라서 에러 pdf의 MMSE 추정값은 0이 된다. 분산은 다음과 같다.

$$\begin{aligned}
 \text{var}(\epsilon) &= \mathbb{E}_{x,\theta}(\epsilon^2) \\
 &= \mathbb{E}_{x,\theta}[(\theta - \hat{\theta})^2]
 \end{aligned} \tag{435}$$

분산은 BMSE의 최소값과 동일하다. 만약 ϵ 가 가우시안 분포를 따르는 경우 이는 다음과 같이 쓸 수 있다.

$$\epsilon \sim \mathcal{N}(0, \text{Bmse}(\hat{\theta})) \tag{436}$$

11.3.1 Example 11.6 - DC Level in WGN - Gaussian Prior PDF

Example 10.1 문제를 다시 생각해보자. 우리는 (365) 추정값을 얻었었다.

$$\hat{A} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \bar{x} + \frac{\frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}} \mu_A \quad (437)$$

그리고 분산은 아래와 같았다.

$$Bmse(\hat{A}) = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \quad (438)$$

에러를 $\epsilon = A - \hat{A}$ 라고 하자. \hat{A} 는 데이터 x 에 선형 관계를 가지므로 x 와 A 는 결합 분포를 이루고(jointly gaussian) 따라서 에러 또한 가우시안 분포를 따른다.

$$\epsilon \sim \mathcal{N}\left(0, \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}}\right) \quad (439)$$

위 식에서 보다시피 $N \rightarrow \infty$ 일수록 에러 pdf는 0에 수렴한다. 충분한 데이터가 관측된 경우 베이지안 추정값은 일관된(consistent) 추정값을 가진다고 하며 \hat{A} 가 충분히 A 에 가까워 졌음을 의미한다.

12 Linear Bayesian Estimators

이전 챕터에서 다뤘던 최적의 베이지안 추정값은 closed form으로 값을 결정하기 어려우며 실제로 구현하기에 계산적으로 매우 복잡하였다. 확률 변수들이 결합 분포를 이룬다는 가정 하에 추정값들은 비교적 쉽게 찾을 수 있었지만 이러한 가정은 실제 추정 문제에서는 참이 아닌 경우가 많다. 이번 챕터에서는 이러한 격차를 메우기 위해 MMSE criterion을 유지하면서도 추정값을 선형으로 제한하는 방법에 대해 소개한다. 그렇게 되면 BLUE에서도 설명하였듯이 추정값은 pdf의 첫 두 모멘트에만 의존하며 closed form 형태로 결정될 수 있다. 여러 방면에서 이 접근법은 고전적인 추정 방법과 유사하다. 실제 추정 문제에서 이러한 추정값들은 위너 필터(Wiener filters)라고 불리며 광범위하게 사용되고 있다.

12.1 Linear MMSE Estimation

스칼라 파라미터 θ 를 추정하기 위해 N 개의 데이터 $x = [x[0], x[1], \dots, x[N-1]]^\top$ 이 관측되었다고 하자. 그리고 베이지안 철학에 의해 θ 또한 확률 변수라고 가정하자. θ 값은 x 의 선형 결합(사실상 affine 결합)으로 다음과 같이 표현할 수 있다.

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] + a_N \quad (440)$$

가중치 계수 a_n 을 찾기 위해 BMSE를 최소화해야 한다.

$$Bmse(\hat{\theta}) = \mathbb{E}[(\theta - \hat{\theta})^2] \quad (441)$$

- $\mathbb{E}[\cdot]$: $p(x, \theta)$ 의 기대값

위 식을 최소화하는 최적의 추정값을 linear MMSE(LMMSE) 추정값이라고 한다. 위 식에서 a_N 이 있다는 것에 주목하자. 이는 x, θ 가 0이 아닌 평균을 가질 때 생성되는 값이다. 만약 두 확률 변수의 평균이 0인 경우 해당 항은 생략된다.

LMMSE 추정값을 결정하기 전에 이는 준최적(suboptimal) 추정값이라는 것을 기억해야 한다. 만약 MMSE 추정값이 선형이 아니라면 이는 최적(optimal) 추정값이 될 수 없다. LMMSE 추정값은 두 확률 변수 x, θ 의 선형 상관 관계에 의존하기 때문에 만약 데이터 x 와 상관 관계가 없거나 비선형 상관 관계인 파라미터 θ 가 존재할 경우 이는 선형 추정값을 통해 예측할 수 없다. 다시 말하면 LMMSE 추정값은 반드시 존재하는 것은 아니다. 다음 예제를 통해 이를 확인해보자.

추정하고자 하는 파라미터 θ 를 단일 관측 데이터 $x[0]$ 로부터 찾고자 한다. 이 때, $x[0] \sim \mathcal{N}(0, \sigma^2)$ 을 따른다고 하자. 추정하고자 하는 파라미터가 $x[0]$ 의 제곱값이라고 하면 최적의 추정값은 아래와 같다.

$$\hat{\theta} = x^2[0] \quad (442)$$

따라서 BMSE 값은 0이 된다. 보다시피 추정값은 비선형 추정값이다. 만약 이 문제에 LMMSE 추정값을 찾는다고 해보자.

$$\hat{\theta} = a_0 x[0] + a_1 \quad (443)$$

두 계수 a_0, a_1 은 BMSE를 최소화함으로써 찾을 수 있다.

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= \mathbb{E}[(\theta - \hat{\theta})^2] \\ &= \mathbb{E}[(\theta - a_0x[0] - a_1)^2] \end{aligned} \quad (444)$$

a_0, a_1 에 대하여 각각 미분하고 0으로 놓으면 다음 식이 유도된다.

$$\begin{aligned} \mathbb{E}[(\theta - a_0x[0] - a_1)x[0]] &= 0 \\ \mathbb{E}[(\theta - a_0x[0] - a_1)] &= 0 \end{aligned} \quad (445)$$

이를 전개하면 다음과 같다.

$$\begin{aligned} a_0\mathbb{E}(x^2[0]) + a_1\mathbb{E}(x[0]) &= \mathbb{E}(\theta x[0]) \\ a_0\mathbb{E}(x[0]) + a_1 &= \mathbb{E}(\theta) \end{aligned} \quad (446)$$

$\mathbb{E}(x[0]) = 0$ 이고 $\mathbb{E}(\theta x[0]) = \mathbb{E}(x^3[0]) = 0$ 므로 계수는 다음과 같다.

$$\begin{aligned} a_0 &= 0 \\ a_1 &= \mathbb{E}(\theta) \\ &= \mathbb{E}(x^2[0]) \\ &= \sigma^2 \end{aligned} \quad (447)$$

따라서 LMMSE 추정값은 $\hat{\theta} = \sigma^2$ 이 되고 이는 데이터 $x[0]$ 에 의존하지 않는다. 이는 θ 와 $x[0]$ 가 서로 상관 관계가 없기(uncorrelated) 때문이다. BMSE 값은 다음과 같다.

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= \mathbb{E}[(\theta - \hat{\theta})^2] \\ &= \mathbb{E}[(\theta - \sigma^2)^2] \\ &= \mathbb{E}[(x^2[0] - \sigma^2)^2] \\ &= \mathbb{E}[x^4[0] - 2\sigma^2\mathbb{E}(x^2[0]) + \sigma^4] \\ &= 3\sigma^4 - 2\sigma^4 + \sigma^4 \\ &= 2\sigma^4 \end{aligned} \quad (448)$$

비선형 추정값 $\hat{\theta} = x^2[0]$ 의 BMSE가 0이었던 것과 대조적으로 LMMSE 추정값은 값을 가진다. 따라서 LMMSE 추정값은 이러한 비선형 문제에 부적절한 것을 알 수 있다.

다시 본론으로 돌아와서 LMMSE에서 최적의 계수 a_n 값을 찾아보자. (440)를 (441)에 넣고 미분을 수행하면 다음과 같은 식이 도출된다.

$$\frac{\partial}{\partial a_N} \mathbb{E}\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n] - a_N \theta\right)^2\right] = -2\mathbb{E}\left[\theta - \sum_{n=0}^{N-1} a_n x[n] - a_N\right] \quad (449)$$

위 식을 0으로 놓고 풀면 다음과 같다.

$$a_N = \mathbb{E}(\theta) - \sum_{n=0}^{N-1} a_n \mathbb{E}(x[n]) \quad (450)$$

앞서 말했듯이 a_N 은 두 확률 변수의 평균이 모두 0이면 a_N 또한 0이 된다. 수식 유도를 계속 이어가서 a_N 을 (441)에 대입해보자.

$$\text{Bmse}(\hat{\theta}) = \mathbb{E}\left\{\left[\sum_{n=0}^{N-1} a_n(x[n] - \mathbb{E}(x[n])) - (\theta - \mathbb{E}(\theta))\right]^2\right\} \quad (451)$$

$\mathbf{a} = [a[0], a[1], \dots, a[N-1]]^\top$ 이라고 했을 때 위 식을 벡터 형태로 바꾸면 다음과 같다.

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= \mathbb{E}\{\left[\mathbf{a}^\top(\mathbf{x} - \mathbb{E}(\mathbf{x})) - (\theta - \mathbb{E}(\theta))\right]^2\} \\ &= \mathbb{E}[\mathbf{a}^\top(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^\top \mathbf{a}] - \mathbb{E}[\mathbf{a}^\top(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\theta - \mathbb{E}(\theta))] \\ &\quad - \mathbb{E}[(\theta - \mathbb{E}(\theta))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^\top \mathbf{a}] + \mathbb{E}[(\theta - \mathbb{E}(\theta))^2] \\ &= \mathbf{a}^\top \mathbf{C}_{xx} \mathbf{a} - \mathbf{a}^\top \mathbf{C}_{x\theta} - \mathbf{C}_{\theta x} \mathbf{a} + \mathbf{C}_{\theta\theta} \end{aligned} \quad (452)$$

- $\mathbf{C}_{xx} \in \mathbb{R}^{N \times N}$
- $\mathbf{C}_{\theta x} \in \mathbb{R}^{1 \times N}$

- $\mathbf{C}_{\theta\theta} \in \mathbb{R}^1$

BMSE를 \mathbf{a} 에 대하여 미분하면 다음과 같다.

$$\frac{\partial \text{Bmse}(\hat{\theta})}{\partial \mathbf{a}} = 2\mathbf{C}_{xx}\mathbf{a} - 2\mathbf{C}_{x\theta} \quad (453)$$

위 식을 0으로 놓고 풀면 최적의 계수 \mathbf{a} 를 구할 수 있다.

$$\mathbf{a} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta} \quad (454)$$

이를 (440)에 대입하면 다음과 같다.

$$\begin{aligned} \hat{\theta} &= \mathbf{a}^\top \mathbf{x} + a_N \\ &= \mathbf{C}_{x\theta}^\top \mathbf{C}_{xx}^{-1} \mathbf{x} + \mathbb{E}(\theta) - \mathbf{C}_{x\theta}^\top \mathbf{C}_{xx}^{-1} \mathbb{E}(\mathbf{x}) \end{aligned} \quad (455)$$

최종적으로 LMMSE 추정값은 다음과 같다.

$$\boxed{\hat{\theta} = \mathbb{E}(\theta) + \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} (\mathbf{x} - \mathbb{E}(\mathbf{x}))} \quad (456)$$

위 식은 MMSE 추정값 (385)과 동일한 형태임에 주목하자. 이는 가우시안 분포에 대한 MMSE 추정값은 선형이 되어 선형 추정값의 조건을 자동으로 만족하기 때문이다. 만약 θ, \mathbf{x} 의 평균이 둘 다 0인 경우 a_N 도 0이 되어 식은 다음과 같아진다.

$$\hat{\theta} = \mathbb{E}(\theta) + \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{x} \quad (457)$$

위 식을 (452)에 넣으면 BMSE는 다음과 같다.

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= \mathbf{C}_{x\theta}^\top \mathbf{C}_{xx}^{-1} \mathbf{C}_{xx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} - \mathbf{C}_{x\theta}^\top \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} - \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} + \mathbf{C}_{\theta\theta} \\ &= \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} - 2\mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} + \mathbf{C}_{\theta\theta} \end{aligned} \quad (458)$$

최종적으로 BMSE 다음과 같다.

$$\boxed{\text{Bmse}(\hat{\theta}) = \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta}} \quad (459)$$

12.1.1 Example 12.1 - DC Level in WGN with Uniform Prior PDF

다음과 같은 데이터 모델이 주어졌다고 하자.

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1 \quad (460)$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN, A 와 독립적(independent)

추정하고자 하는 파라미터 A 는 균일 분포를 따른다고 하자 ($A \sim \mathcal{U}(-A_0, A_0)$). MMSE 추정값은 10.1 섹션에서 보았듯이 적분형이 포함되어 close form으로 구할 수 없었다. 이 문제에 LMMSE 추정값을 사용한다고 하자.

먼저 $\mathbb{E}(A) = 0$ 이므로 $\mathbb{E}(x[n]) = 0$ 을 만족하는 것을 알 수 있다. 벡터 형태로 표현하면 $\mathbb{E}(\mathbf{x}) = 0$ 을 만족한다.

$$\begin{aligned} \mathbf{C}_{xx} &= \mathbb{E}(\mathbf{x}\mathbf{x}^\top) \\ &= \mathbb{E}[(A\mathbf{1} + \mathbf{w})(A\mathbf{1} + \mathbf{w})^\top] \\ &= \mathbb{E}(A^2)\mathbf{1}\mathbf{1}^\top + \sigma^2\mathbf{I} \\ \mathbf{C}_{\theta x} &= \mathbb{E}(A\mathbf{x}^\top) \\ &= \mathbb{E}[A(A\mathbf{1} + \mathbf{w})^\top] \\ &= \mathbb{E}(A^2)\mathbf{1}^\top \end{aligned} \quad (461)$$

- $\mathbf{1} \in \mathbb{R}^{N \times 1}$

(456) 식을 사용하여 LMMSE 추정값을 구하면 다음과 같다.

$$\begin{aligned} \hat{A} &= \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{x} \\ &= \sigma_A^2 \mathbf{1}^\top (\sigma_A^2 \mathbf{1}\mathbf{1}^\top + \sigma^2\mathbf{I})^{-1} \mathbf{x} \end{aligned} \quad (462)$$

- $\sigma_A^2 = \mathbb{E}(A^2)$

위 추정값은 Example 10.2에서 $\mu_A = 0$ 으로 설정했을 때 추정값과 동일하다. 해당 예제에서 추정값을 가져오면 다음과 같다.

$$\hat{A} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \bar{x} \quad (463)$$

$\sigma_A^2 = \mathbb{E}(A^2) = (2A_0)^2/12 = A_0^2/3$ 이므로 이를 대입하여 최종적인 LMMSE 추정값을 구하면 다음과 같다.

$$\hat{A} = \frac{\frac{A_0^2}{3}}{\frac{A_0^2}{3} + \frac{\sigma^2}{N}} \bar{x} \quad (464)$$

적분항이 포함되었던 MMSE 추정값과 달리 LMMSE 추정값은 closed form으로 구할 수 있다. 그리고 식에서 보다시피 A 의 분포에 관계없이 해를 구할 수 있으며 오직 두 개의 모멘트(평균 \bar{x} 과 분산 σ^2)만 알면 된다. 또한 A 와 w 가 독립적이지(independence) 않아도 오직 상관 관계만 없으면(uncorrelated) 된다. 하지만 서두에서 말했듯이 LMMSE는 선형 제약조건이 들어간 suboptimal 추정값이며 MMSE 추정값이 최적 추정값임을 유의하자.

12.2 Geometrical Interpretations

챕터 8에서 least square estimator(LSE)를 기하학적으로 해석하여 벡터 공간에 대한 최소 거리를 찾는 문제임을 설명하였다. LMMSE 또한 이와 유사하게 기하학적으로 해석이 가능하다. 이 때 벡터는 확률 변수가 된다는 점이 챕터 8과 차이점이다. 기하학적 해석은 θ, x 의 평균이 항상 0임을 가정한다. 만약 0이 아니라면 $x' = x - \mathbb{E}(x)$, $\theta' = \theta - \mathbb{E}(\theta)$ 와 같이 0으로 변환하여 θ' 를 추정하는 문제로 정의할 수 있다. LMMSE를 다시 보면 다음과 같다.

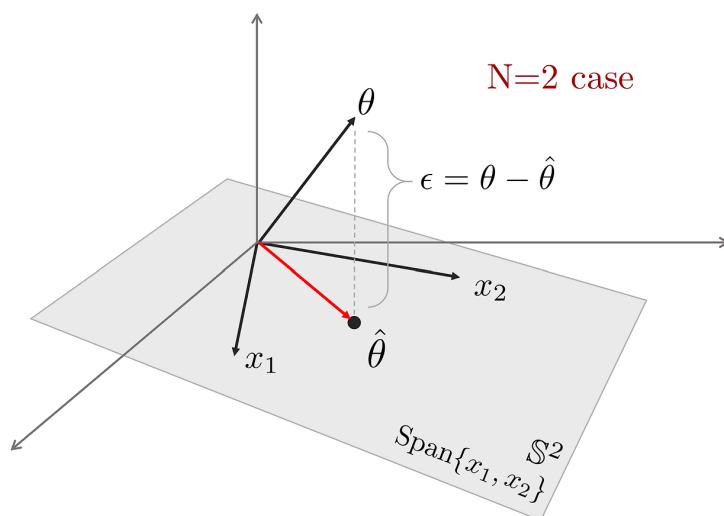
$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] + a_N \quad (465)$$

이는 BMSE를 최소화함으로써 구할 수 있다.

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= \mathbb{E}[(\theta - \hat{\theta})^2] \\ &= \mathbb{E}\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n]\right)^2\right] \\ &= \|\theta - \sum_{n=0}^{N-1} a_n x[n]\|^2 \quad = \|\epsilon\|^2 \end{aligned} \quad (466)$$

위 식을 보면 MSE를 최소화하는 것은 에러 벡터 $\epsilon = \theta - \hat{\theta}$ 의 L2-norm을 최소화하는 것과 동일하다.

LSE와 동일하게 파라미터 θ 는 데이터 벡터 $x = [x[0], x[1], \dots, x[N-1]]^\top$ 이 스펠하는 공간에 존재하지 않으므로 θ 와 x 가 스펠하는 공간의 최소 거리를 찾으면 이는 곧 최적해가 된다. 최소 거리는 평면에 수선의 발을 내림으로써 구할 수 있고 이는 곧 에러 벡터 $\epsilon = \theta - \hat{\theta}$ 가 된다.



따라서 에러 벡터는 데이터 벡터 x 에 직교한다.

$$\epsilon \perp x[0], x[1], \dots, x[N-1] \quad (467)$$

직교성(orthogonality)의 성질을 이용하면 다음과 같은 공식이 유도된다.

$$\boxed{\mathbb{E}[(\theta - \hat{\theta})x[n]] = 0 \quad n = 0, 1, \dots, N-1]} \quad (468)$$

이를 전개하면 다음과 같다.

$$\begin{aligned} \mathbb{E}\left[\left(\theta - \sum_{m=0}^{N-1} a_m x[m]\right)x[n]\right] &= 0 \quad n = 0, 1, \dots, N-1 \\ \text{or} \quad \sum_{m=0}^{N-1} a_m \mathbb{E}(x[m]x[n]) &= \mathbb{E}(\theta x[n]) \quad n = 0, 1, \dots, N-1 \end{aligned} \quad (469)$$

행렬 형태로 표현하면 다음과 같다.

$$\begin{aligned} &\begin{bmatrix} \mathbb{E}(x^2[0]) & \mathbb{E}(x[0]x[1]) & \cdots & \mathbb{E}(x[0]x[N-1]) \\ \mathbb{E}(x[0]x[1]) & \mathbb{E}(x^2[1]) & \cdots & \mathbb{E}(x[1]x[N-1]) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(x[N-1]x[0]) & \mathbb{E}(x[N-1]x[1]) & \cdots & \mathbb{E}(x^2[N-1]) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N - 1 \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}(\theta x[0]) \\ \mathbb{E}(\theta x[1]) \\ \vdots \\ \mathbb{E}(\theta x[N-1]) \end{bmatrix} \end{aligned} \quad (470)$$

이는 정규 방정식 형태이다. 왼쪽 항을 \mathbf{C}_{xx} 라고하고 오른쪽 항을 $\mathbf{C}_{x\theta}$ 라고 하면 위 식은 다음과 같이 나타낼 수 있다.

$$\mathbf{C}_{xx}\mathbf{a} = \mathbf{C}_{x\theta} \quad (471)$$

$$\mathbf{a} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta} \quad (472)$$

따라서 LMMSE 추정값은 다음과 같다.

$$\boxed{\begin{aligned} \hat{\theta} &= \mathbf{a}^\top \mathbf{x} \\ &= \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{x} \end{aligned}} \quad (473)$$

최소 BMSE 값은 \mathbf{a} 를 대입하면 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= \|\epsilon\|^2 \\ &= \|\theta - \sum_{n=0}^{N-1} a_n x[n]\|^2 \\ &= \mathbb{E}\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n]\right)^2\right] \\ &= \mathbb{E}\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n]\right)\left(\theta - \sum_{m=0}^{N-1} a_m x[m]\right)\right] \\ &= \mathbb{E}\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n]\right)\theta\right] - \mathbb{E}\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n]\right)\sum_{m=0}^{N-1} a_m x[m]\right] \\ &= \mathbb{E}(\theta^2) - \underbrace{\sum_{n=0}^{N-1} a_n \mathbb{E}(x[n]\theta)}_{=0} - \underbrace{\sum_{m=0}^{N-1} a_m \mathbb{E}\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n]\right)x[m]\right]}_{=0} \end{aligned} \quad (474)$$

마지막 항은 직교성에 의해 0이 되어 소거된다. 이를 행렬 형태로 표현하면 다음과 같다.

$$\boxed{\begin{aligned} \text{Bmse}(\hat{\theta}) &= \mathbf{C}_{\theta\theta} - \mathbf{a}^\top \mathbf{C}_{x\theta} \\ &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} \end{aligned}} \quad (475)$$

12.3 The Vector LMMSE Estimator

벡터 파라미터에 대한 LMMSE 추정값은 스칼라 파라미터 표현법을 단순히 확장한 것이다. 각각의 파라미터 θ_i 가 모두 BMSE를 최소화하는 추정값이 된다.

$$\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in}x[n] + a_{iN} \quad (476)$$

- $i = 1, 2, \dots, p$

최소화하고자 하는 BMSE는 다음과 같다.

$$Bmse(\hat{\theta}_i) = \mathbb{E}[(\theta_i - \hat{\theta}_i)^2] \quad (477)$$

i 번째 LMMSE 추정값은 다음과 같다.

$$\hat{\theta}_i = \mathbb{E}(\theta_i) + \mathbf{C}_{\theta_i x} \mathbf{C}_{xx}^{-1} (\mathbf{x} - \mathbb{E}(\mathbf{x})) \quad (478)$$

i 번째 최소 BMSE 값은 다음과 같다.

$$Bmse(\hat{\theta}_i) = \mathbf{C}_{\theta_i \theta_i} - \mathbf{C}_{\theta_i x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x \theta_i} \quad (479)$$

스칼라 파라미터 θ_i 는 다음과 같이 벡터 형태로 쓸 수 있다.

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \begin{bmatrix} \mathbb{E}(\theta_1) \\ \mathbb{E}(\theta_2) \\ \vdots \\ \mathbb{E}(\theta_p) \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{\theta_1 x} \mathbf{C}_{xx}^{-1} (\mathbf{x} - \mathbb{E}(\mathbf{x})) \\ \mathbf{C}_{\theta_2 x} \mathbf{C}_{xx}^{-1} (\mathbf{x} - \mathbb{E}(\mathbf{x})) \\ \vdots \\ \mathbf{C}_{\theta_p x} \mathbf{C}_{xx}^{-1} (\mathbf{x} - \mathbb{E}(\mathbf{x})) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}(\theta_1) \\ \mathbb{E}(\theta_2) \\ \vdots \\ \mathbb{E}(\theta_p) \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{\theta_1 x} \\ \mathbf{C}_{\theta_2 x} \\ \vdots \\ \mathbf{C}_{\theta_p x} \end{bmatrix} \mathbf{C}_{xx}^{-1} (\mathbf{x} - \mathbb{E}(\mathbf{x})) \end{aligned} \quad (480)$$

$$\boxed{\hat{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}) + \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} (\mathbf{x} - \mathbb{E}(\mathbf{x}))} \quad (481)$$

- $\mathbf{C}_{\theta x} \in \mathbb{R}^{p \times N}$

BMSE도 벡터 형태로 표현하면 다음과 같다.

$$\begin{aligned} \mathbf{M}_{\hat{\boldsymbol{\theta}}} &= \mathbb{E}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top] \\ &= \mathbf{C}_{\theta \theta} - \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x \theta} \end{aligned} \quad (482)$$

- $Bmse(\hat{\theta}_i) = [\mathbf{M}_{\hat{\boldsymbol{\theta}}}]_{ii}$

이렇듯 LMMSE의 벡터 파라미터 버전은 스칼라 파라미터를 단순 확장한 것이며 기존 LMMSE 성질과 동일하게 pdf의 첫 두 모멘트만 알면 추정값을 구할 수 있다. 마지막으로 LMMSE 추정값의 성질 중 유용한 성질 세 가지를 소개한다.

1. LMMSE 추정값은 선형 변환(또는 affine)에 교환 법칙이 성립한다. 만약 $\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\theta} + \mathbf{b}$ 와 같은 선형 변환된 파라미터 $\boldsymbol{\alpha}$ 를 추정하고자 할 때 LMMSE 추정값은 $\hat{\boldsymbol{\alpha}} = \mathbf{A}\hat{\boldsymbol{\theta}} + \mathbf{b}$ 이 된다.
2. 만약 서로 독립인 두 파라미터 θ_1, θ_2 가 주어졌을 때 $\boldsymbol{\alpha} = \theta_1 + \theta_2$ 를 만족하는 $\boldsymbol{\alpha}$ 를 추정하고자 하는 경우 이는 $\hat{\boldsymbol{\alpha}} = \hat{\theta}_1 + \hat{\theta}_2$ 와 같이 쉽게 추정값을 찾을 수 있다.
3. LMMSE 추정값은 MMSE 추정값이 선형이며 가우시안 분포를 따르는 경우와 동일한 추정값을 제공한다.

12.3.1 Theorem 12.1 (Bayesian Gauss-Markov Theorem)

만약 데이터가 베이지안 선형 모델로 다음과 같이 구성되어 있는 경우

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \quad (483)$$

- $\mathbf{x} \in \mathbb{R}^{N \times 1}$
- $\mathbf{H} \in \mathbb{R}^{N \times p}$
- $\boldsymbol{\theta} \in \mathbb{R}^{p \times 1}$
- $\mathbf{w} \sim \mathcal{N}(0, \mathbf{C}_w)$: WGN, $\boldsymbol{\theta}$ 와 상관 관계 없음(uncorrelated)

LMMSE 추정값은 다음과 같다.

$$\boxed{\begin{aligned} \hat{\boldsymbol{\theta}} &= \mathbb{E}(\boldsymbol{\theta}) + \mathbf{C}_{\theta\theta} \mathbf{H}^\top (\mathbf{H} \mathbf{C}_{\theta\theta} \mathbf{H}^\top + \mathbf{C}_w)^{-1} (\mathbf{x} - \mathbf{H} \mathbb{E}(\boldsymbol{\theta})) \\ &= \mathbb{E}(\boldsymbol{\theta}) + (\mathbf{C}_{\theta\theta}^{-1} + \mathbf{H} \mathbf{C}_w^{-1} \mathbf{H}^\top)^{-1} \mathbf{H}^\top \mathbf{C}_w^{-1} (\mathbf{x} - \mathbf{H} \mathbb{E}(\boldsymbol{\theta})) \end{aligned}} \quad (484)$$

추정값의 성능은 에러 벡터 $\boldsymbol{\epsilon} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ 를 평가함으로써 알 수 있다. 에러 벡터는 $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{C}_\epsilon)$ 을 따른다.

$$\begin{aligned} \mathbf{C}_\epsilon &= \mathbb{E}_{x,\theta}(\boldsymbol{\epsilon} \boldsymbol{\epsilon}^\top) \\ &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta\theta} \mathbf{H}^\top (\mathbf{H} \mathbf{C}_{\theta\theta} \mathbf{H}^\top + \mathbf{C}_w)^{-1} \mathbf{H} \mathbf{C}_{\theta\theta} \\ &= (\mathbf{C}_{\theta\theta}^{-1} + \mathbf{H}^\top \mathbf{C}_w^{-1} \mathbf{H}^\top)^{-1} \end{aligned} \quad (485)$$

에러 벡터의 공분산 \mathbf{C}_ϵ 은 또한 최소 BMSE 값을 의미한다.

$$\begin{aligned} [\mathbf{M}_{\hat{\boldsymbol{\theta}}}]_{ii} &= [\mathbf{C}_\epsilon]_{ii} \\ &= \text{Bmse}(\hat{\theta}_i) \end{aligned} \quad (486)$$

13 Kalman Filters

14 References

- [1] Kay, Steven M. Fundamentals of statistical signal processing: estimation theory. Prentice-Hall, Inc., 1993.
- [2] Simon, Dan. Optimal state estimation: Kalman, H infinity, and nonlinear approaches. John Wiley Sons, 2006.

15 Revision log

- | | | |
|-------------------|-------------------|-------------------|
| • 1st: 2024-02-09 | • 3rd: 2024-02-11 | • 5th: 2024-02-18 |
| • 2nd: 2024-02-10 | • 4th: 2024-02-13 | • 6th: 2024-02-19 |